Transitioning from Fixed-Length Questionnaires to Computer-Adaptive Versions

Otto B. Walter and Heinz Holling

University of Münster, Germany

Abstract. To investigate how an existing questionnaire can be transformed into a computer-adaptive version, we developed an adaptive version of the Interpersonal Competence Questionnaire (ICQ). This adaptive version was based on a representative sample ($N = 1934$) of respondents who answered 30 items from a German translation of the ICQ. A random half of the sample was used to evaluate test dimensionality, calibrate the items, and model the relation between person parameters and raw total scores. The other random half of the sample was employed to assess the comparability of person parameters and raw scores. After these tests and item calibration, 28 items remained in the item pool. A high correlation was found between raw scores and estimated scores using all items. Raw scores could be predicted accurately from estimated person parameters. These results indicate that our approach is an effective technique for transforming an existing questionnaire into a computer-adaptive version.

Keywords: item response theory, computer-adaptive test, social competence, questionnaire

Introduction

Computer-adaptive tests (CATs) were initially developed for ability and achievement testing in large-scale educational assessments. A particularly attractive property of CATs in this context is the possibility of assembling tests tailored to the *individual* ability level of each respondent. The test score, however, can be computed on a *common* metric, even though different respondents may have answered different sets of items. Over the last decade, a rising interest in the development of CATs outside educational testing can be observed. A prominent example of this is the joint initiative working on building the "Patient-Reported Outcomes Measurement Information System" sponsored by the U.S. National Institutes of Health (NIH) (Cella et al., 2007).

There are some differences in the application of CATs for the measurement of psychological, social, or health-related constructs in comparison to CATs in educational assessment. Ability and achievement tests most frequently use binary scored (correct/wrong) multiple choice items that can be analyzed by dichotomous item response theory (IRT) models. In contrast, personality traits and social constructs are usually assessed by items that are scored on a Likert-type scale. For the analysis of such items, polytomous IRT models have been proposed. These models provide high item information over a large range of the latent trait. As the asymptotic standard error of a response pattern is the reciprocal of the square root of the sum of item information of this response pattern, a CAT algorithm that terminates when a given error bond is reached will terminate sooner if item information is higher. Therefore, the use of polytomous IRT models opens up the opportunity for substantial item savings and may help to reduce the item burden placed on respondents. Notwithstanding the advantages that the use of CATs in applied psychological measurement promises, several problems remain to be solved. One of these open questions deals with the transformation of existing fixed-length questionnaires into adaptive tests. There is, to date, no golden standard for constructing CATs. Moreover, it would be beneficial if more detail could be given in respect to the comparability of scores obtained from the computer-adaptive version and a paper and pencil version of a questionnaire. This point seems to be of particular importance, as the IRT framework requires that a number of rather strong assumptions be met (e.g., unidimensionality of the underlying construct), which may affect the construct that is measured by the adaptive test.

An example for the challenges that occur when an existing questionnaire is transformed into a unidimensional computer-adaptive test can be found in the assessment of *interpersonal competence* as measured by the Interpersonal Competence Questionnaire (ICQ; Buhrmester, Furman, Wittenberg, & Reis, 1988), in which five dimensions of social competence are assessed. The questionnaire has been applied in several studies within the framework of developmental and clinical psychology (Buhrmester et al., 1988; Buhrmester, 1990; Lamke, Sollie, Durbin, & Fitzpatrick, 1994; Schneider & Younger, 1996; Semple, Shaw, Grant, Moscona, & Jeste, 1999; Miller & deWinstanley, 2002).

DOI 10.1027/0044-3409.216.1.22

Based on the framework of IRT, we developed an item bank and a CAT algorithm that can measure *interpersonal competence* using a computerized adaptive method. The construction of the item bank was based on data obtained previously from a poll of a representative German sample, in which a German version of the Interpersonal Competence Questionnaire (Buhrmester et al., 1988; Riemann & Allgöwer, 1993; Kanning, 2006) was administered.

An item bank of a computer-adaptive version of the ICQ is supposed to include only those items that meet the statistical requirements of the underlying IRT framework. The purpose of our study was to develop a CAT to measure *interpersonal competence* that can be used as an alternative to the fixed-length version of the ICQ. Our aim was to evaluate the extent to which an existing questionnaire can be converted into a computer-adaptive version. This article outlines the steps of the development of the ICQ-CAT and describes its properties in comparison to the fixed-length version using simulation studies based on real data. We placed particular focus on the comparability of raw scores obtained from the fixed-length questionnaire and the computer-adaptive version.

Methods

In this section, we will outline the development of a computer-adaptive version of the fixed-length ICQ. We describe the German version of the ICQ employed in this study, present the samples that were used to calibrate the item bank, and conduct the simulation studies. We then summarize the computer-adaptive algorithm and outline the derivation of a prediction model of raw scores from CAT scores.

German Version of the ICQ

The ICQ (40 items) was translated into German by Riemann and Allgöwer (1993). Kanning and Holling (1999) proposed a slightly modified version of this translation. To reduce central tendency bias, they suggested that the respondents' agreement be indicated on a 4-point scale rather than on the original 5-point scale. Using psychometric properties computed from a sample of $N = 1955$ policemen and policewomen (Kanning, 2006), they also developed a shortened version (30 items). This shortened version of the ICQ was used in our present study.

Data Sample

Data were obtained in 2003 from a representative German polling of 2089 respondents who answered, among other questionnaires, the German version of the ICQ described above. For our analysis, we considered only the complete data sets from 1934 respondents. Exclusion of respondents with incomplete data sets did not yield any significant changes in the age or in the sex distribution of the sample, and it was therefore assumed that these missing data had little, if any, impact on the analysis presented here. The sample was comprised of 49.5% male and 50.5% female respondents. The age of the respondents ranged between 14 years and 90 years, with a mean of 48 years. The sample was split randomly into two nonoverlapping halves of 967 respondents each. The first half of the sample (Sample A) was employed to develop an item bank of the computer adaptive version of the ICQ. The properties of the resulting item bank were investigated in simulation studies using the responses of the second half of the sample (Sample B) only.

Empirical Item Analysis

The item bank of the computer-adaptive version of the ICQ was constructed following the procedure described in detail elsewhere (Walter et al., 2005, 2007; Fliege et al., 2005). Employing a similar approach as the one proposed by Ware, Bjorner, and Kosinski (2000), we examined whether the items were measuring one underlying dimension, conducted a visual inspection of item response functions (nonparametric analysis of IRFs), and calibrated the items using a polytomous IRT model.

Assessment of Unidimensionality and Local Dependence

The question as to whether the items are measuring one underlying dimension or separate dimensions is a crucial issue in IRT. Various methods have been suggested for determining unidimensionality (for a review, see Hattie, 1984, 1985). Thissen, Reeve, Bjorner, and Chang (2007) point out that there are no definitive rules for deciding when multidimensionality or local dependence is of significant magnitude to cause problems. We followed the approach proposed by Bjorner, Kosinski, and Ware (2003), who investigated residual correlations after fitting a one-factor model using polychoric correlations. This approach has also proved to be a useful criterion during the development of polytomous CATs for anxiety and depression (Walter et al., 2005, 2007; Fliege et al. 2005).

Choice of IRT Model

Over the past few decades, a number of parametric unidimensional IRT models have been proposed (Thissen & Steinberg, 1986; van der Linden & Hambleton, 1997). The Partial Credit Model (PCM; Masters, 1982; Masters & Wright, 1997), the Generalized Partial Credit Model (GPCM; Muraki, 1992, 1997), and the Graded Response

Model (GRM; Samejima, 1969, 1997) are appropriate models for polytomous and ordered item responses. A distinguishing feature of these models is whether the item discrimination parameter (slope) is set to vary across items (GPCM, GRM) or is a common characteristic for all items (PCM). Masters and Wright (1997) argue that the strength of the PCM is its parsimony and the fact that it shares the statistical properties of the IRT models of the Rasch family, such as sufficient statistics for all model parameters and separable person and item parameters. However, it has been noted that a less constrained model that estimates separate slopes for each item can often provide a more accurate reflection of the data (Edelen & Reeve, 2007). Visual inspection of item response functions computed nonparametrically may be helpful in determining which class of models is appropriate (Bjorner et al., 2003; Edelen & Reeve, 2007; Bjorner, Chang, Thissen, & Reeve, 2007). Edelen and Reeve (2007), as well as Bjorner et al. (2007), noted that the choice between the GPCM and the GRM is somewhat arbitrary, as these two models generally produce nearly identical results.

Visual analysis of item response functions suggested that the steepness of these functions varies considerably between the items. Therefore, the GPCM was chosen to calibrate the items. This model can be characterized as follows. Let X_{ij} denote the response of person *j* to item *i* given ability θ_j. The GPCM assumes that the probability of choosing category *h* over category *h*–1 is governed by the logistic dichotomous response model:

$$\ln\frac{P(X_{ij}=h|\theta_j)}{P(X_{ij}=h-1|\theta_j)}=\alpha_i(\theta_j-b_{ic}),\ h>0,\ b_{i0}\equiv 0$$

CAT Algorithm

The specification of a computer adaptive algorithm requires (1) a rule for selecting items to be presented to the respondent (item selection rule), (2) a procedure for determining the current estimate of the latent trait (latent trait estimation), and (3) a rule for deciding whether the algorithm should terminate or continue with selecting items (stopping rule).

Van Rijn, Eggen, Hemker, and Sanders (2002) distinguish between two main approaches currently being used to select items. The first approach is based on item information. This rule selects the most informative item at the current estimated ability level. The most commonly used form of item information is Fisher information, but Kullback-Leibler information has also been studied in CATs (e.g., Chang & Ying, 1996; Eggen, 1999). The second approach is Bayesian item selection based on a prior or posterior distribution of ability and a Bayesian variant of item information (van der Linden, 1998). Van der Linden and Pashley (2000) note that even though no asymptotic motivation existed for the use of the maximum information as item selection criterion in CATs, this criterion immediately became a popular choice in adaptive testing. The popularity of this criterion may be due to its easy implementation and, as several simulation studies indicate (van der Linden & Pashley, 2000; van Rijn et al., 2002; Veldkamp, 2003), due to the fact that the bias introduced by several item selection criteria becomes small as the number of items administered increases. In this study, we decided to use maximum Fisher information as item selection criterion because of its widespread use in operating CATs with polytomous items (e.g., Ware et al., 2000; Bjorner et al., 2003; Fliege et al., 2005; Bjorner et al., 2007; Walter et al., 2007).

Similar practical considerations led us to choose *expected a posteriori estimation* (Bock & Aitken, 1981) for latent trait estimation. It has been noted that person parameter estimates using this approach may be biased toward the prior mean (Chen, Hou, & Dodd, 1998; Meijer & Nering, 1999). However, results from simulation studies indicate that the bias is small for ability levels around two standard deviations above or below the mean (Bock & Mislevy, 1982; Wang & Hanson, 1999). One of the major advantages of EAP estimation is that the estimate can be computed easily using a noniterative procedure. In contrast to maximum likelihood estimation, the procedure provides a finite estimate even in cases with extreme response patterns. The estimation starts with an assumption about the distribution φ of the ability θ in the population (e.g., φ is the standard normal density). For a vector of responses h_i of length N, the item response functions of the chosen categories h_i are multiplied with the prior distribution. The IRT score estimate can then be computed as the mean of the resulting posterior distribution:

$$\hat{\theta}_j=\frac{\int_{-\infty}^{+\infty}\theta\varphi(\theta)\prod_{i=1}^{N}P(X_{ij}=h_i\,|\,\theta)\,\mathrm{d}\theta}{\int_{-\infty}^{+\infty}\varphi(\theta)\prod_{i=1}^{N}P(X_{ij}=h_i\,|\,\theta)\,\mathrm{d}\theta}$$

These integrals can be solved by a numeric procedure as described by Bock and Mislevy (1982), who also provide an equation to compute the posterior standard deviation (PSD) of the posterior. The PSD plays the same role as the asymptotic standard error of the maximum likelihood estimator.

The CAT algorithm we implemented to estimate the latent trait can be summarized as follows. Initially, the estimate is set to the assumed population mean ($\theta_0 = 0.0$). For this estimate, the item with the highest Fisher information is selected and presented to the respondent. After the respondent's answer is recorded, EAP estimation as described above is used to compute an ability estimate and PSD. The computed ability estimate determines the next item via the maximum information rule, which is then presented to the respondent. These steps are repeated until either the PSD falls below 0.32 (stopping rule) or all items have been presented. Since we assume that the ability has a standard normal distribution, the intraclass correlation $\rho = 1 - [\mathrm{PSD}(\theta)]^2$ is the reliability coefficient for the EAP estimate (Bock & Mislevy, 1982). Thus, the criterion

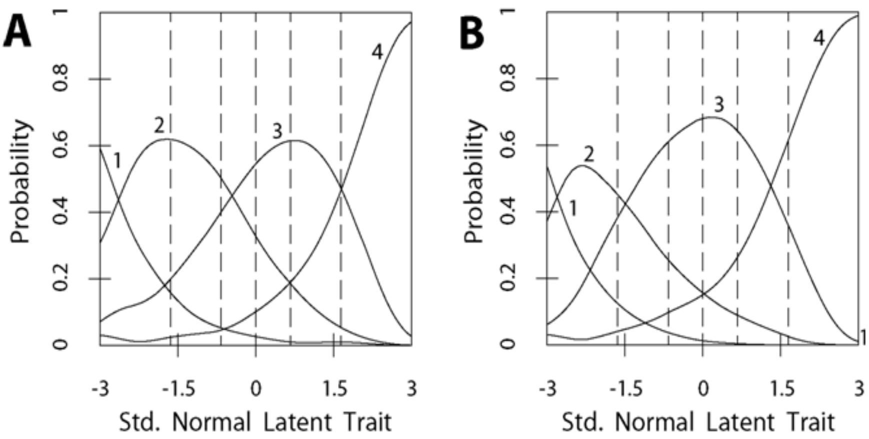

Figure 1. Nonparametrically computed item response functions for standard normal scores of the ICQ-CAT. (A) Item text: “Telling a companion you don’t like a certain way he or she has been treating you.” (B) Item text: “Helping a close friend through his or her thoughts and feelings about a major life decision, e.g., a career choice.”

PSD ≤ 0.32 can be interpreted as corresponding to a reliability of ρ ≥ 0.9. The test result is defined by the estimate and PSD computed in the step before the algorithm was terminated.

Prediction of Raw Total Scores from CAT Scores

Using our CAT algorithm, we estimated the latent trait for all 934 respondents of Sample A. For this estimation, all available items in the item pool were used. The actual score of the respondent was then predicted by a linear quadratic regression model from the estimated ability levels. This choice of a prediction model was motivated by our intention to examine whether a simple procedure which can be used in practice can provide a reasonably accurate conversion between IRT and raw total scores. Moreover, an accurate conversion between these scores can serve as an empirical check on the IRT model. It is worth noting that (except for a middle section of the test-characteristic function) IRT scores and total raw scores are on a different scale. The accuracy of the prediction was investigated in several simulation studies.

Simulation Studies

The simulation studies were conducted with the responses from Sample B. As mentioned earlier, this sample did not overlap with Sample A, which was used to calibrate the items of the ICQ-CAT. However, responses to all items of the ICQ-CAT were available in Sample B. It should be noted that such simulations are based on the assumption that the order of item presentation and the number of items administered have only little impact on the estimation of the person parameter (Gardner, Kelleher, & Pajer, 2002), which is as expected under the IRT model.

In the first study, ability levels were estimated by simulated CAT runs on the basis of the responses from Sample B. These estimates were correlated with the scores resulting from the use of all items to estimate the latent trait. This correlation can be seen as a measure of the information loss due to the restricted number of items that are presented in a CAT. Furthermore, correlations between CAT scores and ICQ raw scores were computed for this sample.

A second study was conducted to evaluate the accuracy of the prediction of ICQ raw scores by the quadratic regression equation derived from Sample A. Using the scores from the simulated CAT runs, we predicted raw scores for the respondents of Sample B from our regression model. We computed the correlation between ICQ raw scores and predicted scores, and examined the distribution of the prediction errors (i.e., the difference between predicted scores and ICQ raw scores).

Results

Item Bank Construction

Unidimensionality

To determine the extent to which items are unidimensional, we conducted a one-factorial factor analysis for categorical variables using MPlus (Muthén & Muthén, 2004). Two items were excluded to avoid any residual correlations larger than .25 (“Saying ‘no’ when a friend asks you to do something you don’t want to do,” and “Not exploding at a close companion in order to avoid a damaging conflict.”). A subsequent confirmatory factor analysis yielded no residual correlations larger than .25.

Investigation of Item Response Curves

Visual inspection of item response functions (IRFs) computed nonparametrically (Gaussian kernel smoothing; Ramsay, 1995) as proposed by Bjorner et al. (2003) proved to be a useful step during the analysis. In this step, the shapes of observed IRFs are compared to the shapes of parametrically modeled response functions. Ideally, an item exhibits peaked response functions that exceed all other functions over exactly one interval of the latent trait. Furthermore, the values for which an item is maximal should match the order in which the response options of an item are presented. These criteria were met by all items. Figure 1 shows two example items.

Item Calibration

For item calibration using the GPCM, the metric was set in reference to a population mean of 0 and a standard deviation of 1. Item parameters were estimated by a marginal likelihood estimation procedure provided by the *Parscale* software. Item information and, in turn, measurement precision is determined, to a great extent, by the slope parameter. After item calibration, all but one item exhibited sufficiently high slope parameters ranging between 0.59 and 1.46 (mean ± *SD*: 1.04 ± 0.2). The final item pool of the ICQ-CAT comprised 28 items and was investigated in several simulation studies.

Properties of the Item Bank

Ability levels for all respondents of Sample B were estimated by simulated CAT runs in which the stopping rule was set to PSD ≤ .32. The average test length necessary to realize this precision across the test administrations was 17.1 ± 3.4 items (mean ± *SD*). In a second set of simulated CAT runs, all 28 items were used to estimate the person parameters of Sample B. The correlation between CAT scores obtained when all items were administered and when the stopping of PSD ≤ 0.32 was applied was very high (r = .97, see Figure 2). This indicates that, despite substantial item savings of about 40% due to the CAT algorithm, not much information is lost and a precise estimation of the latent trait is still possible.

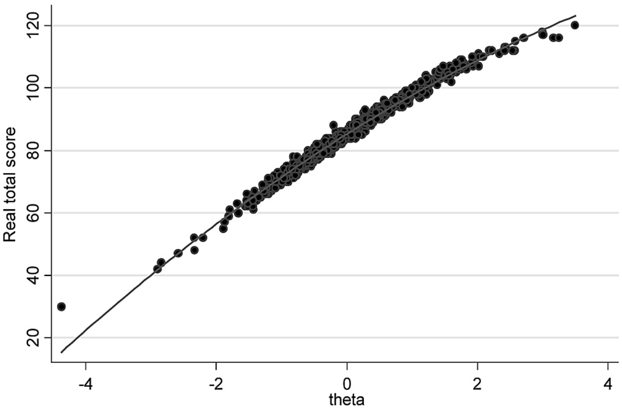

Figure 2. ICQ total scores (30 items) as a function of ability estimates obtained from all 28 items of the item pool in calibration Sample A. Note: Quadratic linear regression was R^2 = 0.98.

Prediction of Raw Scores from CAT Scores

A set of simulated CAT runs in which all 28 items were administered was also conducted for the calibration Sample A. The correlation between estimated IRT scores (28 items) and ICQ raw scores (30 items) was very high (r = .98) and the relation between these scores can be modeled closely by a quadratic regression function (R^2 = 0.98, see Figure 3).

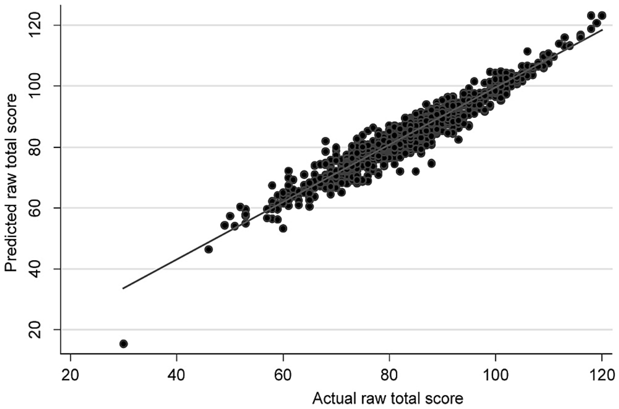

Figure 3. Predicted raw total scores as a function of the actual total scores of the ICQ in Sample B.

Using the simulated CAT scores in which the stopping rule was set to PSD ≤ 0.32, we predicted the raw total scores of all respondents of Sample B using the quadratic regression model. Even though Sample B was neither used in the construction of the item bank nor in the modeling of the regression function, we found a close relation between CAT scores and total scores. The linear regression line between these two variables nearly coincided with the identity line (R^2 = .93). An examination of the distribution of the prediction errors (defined here as the difference between predicted scores and actual raw scores) showed a narrow distribution around zero (mean ± *SD*: –0.38 ± 3.31). 95% of all deviations between predicted and real raw score are contained in the interval [–6.6; 6.4]. Considering that the raw total scores in the sample range between 30 and 120 (mean ± *SD*: 84.3 ± 12.5), this indicates that scores from CAT runs can be accurately converted into raw scores and vice versa.

Discussion

The ICQ-CAT presented here was constructed from an existing fixed-length questionnaire. There are substantial differences between tests developed within the framework of item response theory and classical test theory. Most IRT models require a unidimensional underlying latent trait and Hambleton, Swaminathan, and Rogers (1991) consider unidimensionality one of the most important aspects in determining the quality of the IRT model. In contrast, most standardized questionnaires exhibit a factorial structure corresponding to several scales. These scales may or may not be highly correlated. If the correlation between scales is low, it is likely that these scales violate the unidimensionality needed by the IRT framework. However, the ICQ is an example of a questionnaire with high correlations between its five scales (Buhrmester et al., 1988; Kanning, 2006). These high correlations may indicate that the five dimensions really represent facets of a higher order construct (*social competence*). In this respect, the investigation of residual correlations after fitting a one-factorial model for categorical variables proved to be a valuable criterion during the construction of the item bank. The cut-off level of 0.25 appears to be flexible enough both to tolerate the existence of several scales and to capture a higher order construct. The particular choice of a cut-off level of 0.25 was motivated by reports that item calibration is to some extent robust to slight violations of unidimensionality (Drasgow & Parsons, 1983; Reckase, 1979), and by Bjorner et al. (2003), who employed a similar, slightly more conservative cut-off of 0.20. The simulation studies showed that scores obtained from the ICQ-CAT, even in its present form, can accurately predict the raw total score. The possibility to convert CAT scores into raw total scores and vice versa is indispensable for comparisons between an adaptive and the fixed-length versions of the instrument.

The construction of the ICQ-CAT is based on positive experiences gained in the development of CATs measuring anxiety and depression, which also use polytomous items and aim at capturing constructs not related to achievement and ability assessment. These CATs not only allow for a precise measurement of psychological constructs but may also help to reduce the test burden placed on respondents.

Even though the ICQ-CAT shows the advantages that are expected from theory and captures the underlying construct in a similar manner as the original fixed-length questionnaire, there are areas for possible refinements. So far, item selection is determined solely by a statistical criterion (maximum item information). To improve comparisons across the adaptive and the fixed-length version, it would be advantageous if items were selected not only on item information, but also on content, preferably on all five scales. Notwithstanding the current limitations, the approach presented here for transforming an existing questionnaire into a computer-adaptive version has the potential to broaden our range of test administration options in applied psychological measurement.

References

Bjorner, B.B., Chang, C.H., Thissen, D., & Reeve, B.B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research, 16,* 95–108.

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., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

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

Buhrmester, D. (1990). Intimacy of friendship, interpersonal competence, and adjustment during preadolescence and adolescence. *Child Development, 61,* 1101–1111.

Buhrmester, D., Furman, W., Wittenberg, M.T., & Reis, H.T. (1988). Five domains of interpersonal competence in peer relationships. *Journal of Personality and Social Psychology*, 55, 991–1008.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B. et al. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH Roadmap Cooperative Group during its first two years. *Medical Care, 45,* I3-I11.

Chang, H.H., & Ying Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213–229.

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

Drasgow, F., & Parsons, C.K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189–199.

Edelen, M.O., & Reeve, B.B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16,* 5–18.

Eggen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249–261.

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.

Gardner, W., Kelleher, K.J., & Pajer, K.A. (2002). Multidimensional adaptive testing for mental health problems in primary care. *Medical Care, 40,* 812–823.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19,* 49–78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9,* 139–164.

Kanning, U.P. (2006). Development and validation of a German-language version of the Interpersonal Competence Questionnaire (ICQ). *European Journal of Psychological Assessment, 22,* 43–51

Kanning, U.P., & Holling, H. (1999). *Testtheoretische Analysen zum Interpersonalen Kompetenzfragebogen (ICQ) in einer realen Auswahlsituation* [Test theoretic analysis of the Interpersonal Competence Questionnaire (ICQ) in a reality-based selection procedure]. Unpublished manuscript, Universität Münster, Germany.

Lamke, L.K., Sollie, D.L., Durbin, R.G., & Fitzpatrick, J.A.

(1994). Masculinity, femininity and relationship satisfaction: The mediating role of interpersonal competence. *Journal of Social & Personal Relationships, 11,* 535–554.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.

Masters, G.N., & Wright, B.D. (1997). The partial credit model. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). Berlin: Springer-Verlag.

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

Miller, J.B., & deWinstanley, P.A. (2002). The role of interpersonal competence in memory for conversation. *Personality and Social Psychology Bulletin, 28,* 78–89.

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

Muraki, E. (1997). A generalized partial credit model. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). Berlin: Springer-Verlag.

Muthén, L.K., & Muthén, B.O. (2004). *Mplus. The comprehensive modeling program for applied researchers;. Users guide* [Computer software and manual]. Los Angeles: Muthén & Muthén.

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

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Riemann, R., & Allgöwer, A. (1993). Eine deutschsprachige Fassung des Interpersonal Competence Questionnaire (ICQ) [A German version of the Interpersonal Competence Questionnaire (ICQ)]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14,* 153–163.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34,* 110–114.

Samejima, F. (1997). The graded response model. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). Berlin: Springer-Verlag.

Schneider, B.H., & Younger, A.J. (1996). Adolescent-parent attachment and adolescents' relations with their peers: A closer look. *Youth and Society, 28,* 95–108.

Semple, S.J., Shaw, W.S., Grant, I., Moscona, S., & Jeste, D.V. (1999). Self-perceived interpersonal competence in older schizophrenia patients: The role of patient characteristics and psychosocial factors. *Acta Psychiatrica Scandinavica, 100,* 126–135.

Thissen, D., Reeve, B.B., Bjorner, J.B., & Chang, C.H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research, 16,* 109–119.

Thissen, D., & Steinberg, L. (1986). Taxonomy of item response models. *Psychometrika, 51,* 567–577.

van der Linden, W.J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63,* 201–216.

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

van der Linden, W.J., & Pashley, P.J. (2002). Item selection and ability estimation in adaptive testing. In W.J. van der Linden, & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Dordrecht, The Netherlands: Kluwer.

van Rijn, P.W., Eggen, T.H.J.M., Hemker, B.T., & Sanders, P.F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26,* 393–411.

Veldkamp, B.P. (2003). Item selection in polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.), *New developments in psychometrics* (pp. 207–214). Tokyo: Springer-Verlag.

Walter, O.B., Becker, J., Bjorner, J.B., Fliege, H., Klapp, B.F., & Rose M. (2007). Development and evaluation of 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. et al. (2005). Entwicklungsschritte für einen computeradaptiven Test zur Erfassung von Angst (A-CAT) [Development of a computer adaptive test for anxiety (A-CAT)]. *Diagnostica, 51,* 88–100.

Wang, T., & Hanson, B.A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23,* 263–278.

Ware, J.E., Bjorner, J.B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing. A brief summary of ongoing studies of widely used headache impact scales. *Medical Care, 38,* II73–II82.

Otto B. Walter

Lehrstuhl für Statistik und Methoden
Fachbereich Psychologie und Sportwissenschaft
Psychologisches Institut IV
Westfälische Wilhelms-Universität Münster
Fliednerstr. 21
D-48149 Münster
Germany
Tel. +49 251 8339-140
Fax +49 251 8339-419
E-mail otto.walter@uni-muenster.de