Uncertainties in the item parameter estimates and robust automated test assembly

Bernard P. Veldkamp, University of Twente, The Netherlands

Mariagiulia Matteucci, University of Bologna, Italy

Martijn de Jong, Erasmus University Rotterdam, The Netherlands

Abstract

IRT parameters have to be estimated, and because of the estimation process, they do have uncertainty in them. In most large scale testing programs, the parameters are stored in item banks, and automated test assembly algorithms are applied to assemble operational test forms. These algorithms treat item parameters as fixed values, and uncertainty is not taken into account. As a consequence, resulting tests might be off target or less informative than expected. In this paper, the process of parameter estimation is described to provide insight into the causes of uncertainty in the item parameters. The consequences of uncertainty are studied. Besides, an alternative automated test assembly algorithm is presented that is robust against uncertainties in the data. Several numerical examples demonstrate the performance of the robust test assembly algorithm, and illustrate the dangers of not taking this uncertainty into account. Finally, some recommendations about the use of robust test assembly and some directions for further research are given.

Keywords: Automated test assembly, Computerized adaptive testing, Item parameter estimation, Item response theory, Robust optimization, Robust test assembly, Uncertainty.

Introduction

In most large scale testing programs, whether they use computerized adaptive tests or not, items are selected from an item bank to assemble operational test forms. Many methods have been developed for automated test assembly. For paper-and-pencil (P&P) tests, 0-1 linear programming techniques are commonly applied. They are either based on network programming (Armstrong, Jones, Wang, 1995), integer programming (van der Linden, 2005), or on heuristic methods (Belov, & Armstrong, 2005; Luecht, & Hirsch, 1992; Stocking, and Swanson, 1993, Veldkamp, 2002, Verschoor, 2007). For computer adaptive testing (CAT), linear programming techniques or heuristics are generally combined with exposure control methods (Barrada, Abad, & Veldkamp, 2009; Barrada, Olea, & Ponsoda, 2007; Barrada, Veldkamp, & Olea, 2009; Chang & Ying, 1999; Revuelta, & Ponsoda, 1998; Sympson, & Hetter, 1985; van der Linden, & Veldkamp, 2004, 2007).

Even though these automated test assembly methods differ with respect to many aspects, they do have one thing in common. All methods assume that item parameters can be treated as fixed values without any uncertainty in them. Obviously, this assumption can never be met. Item parameters have to be estimated. They cannot be observed directly. Due to the estimation process, some error of measurement is involved. To provide more insight in the estimation process, a detailed description of the most common estimation methods in IRT is provided in the next section.

Both uncertainties in the amount of information the items provide, and in the location where the items provide maximum information affect the test assembly process.
For example, when items are selected based on maximum information, items with over-estimated discrimination parameters tend to be favored, since these items provide most information. Item selection based on maximum information, therefore, capitalizes on positive estimation errors if this uncertainty is not taken into account. An alternative approach would be to include uncertainties due to the estimation errors in the item selection process. Items can be stored in the item bank in such way that uncertainties are stored as well, and automated test assembly procedures could use this information.
Robust optimization techniques have been proposed to deal with optimization problems with uncertainties in the parameters. The question remains whether they are applicable to the problem of automated test assembly.
The focus of this paper is on the problem of how to deal with uncertainty in the item parameters in the test assembly process. We would like to answer the following questions. Why do we have uncertainty in test assembly problems? How serious is the problem? Would it be an option to apply robust optimization techniques to automated test assembly, in order to deal with uncertainties involved?

Uncertainty in item parameter estimation

To answer these questions, we first have to address how uncertainty plays a role in item parameter estimation more into detail. In educational and psychological measurement, test theory, e.g. item response theory (IRT), is applied to determine a probabilistic

relationship between the responses of a candidate and his/her underlying ability. IRT was outlined in the sixties (Lord and Novick, 1968) but developments in research are still very active (De Ayala, 2009; Reckase, 2009). In IRT models, uni-dimensionality and local independence are assumed. Besides, a number of conditions are imposed on the item response function (IRF) describing the mathematical relation between the probability of success $P_i(\theta)$ and the underlying ability θ. A logistic (Birnbaum, 1968; Rasch, 1960), or a cumulative normal distribution function (Lord, 1952) is most commonly applied. Parameters describing the IRFs, also called the item parameters, are estimated based on test responses. Both maximum likelihood and Bayesian estimation methods can be used. Parameters of logistic models are commonly estimated within marginal maximum likelihood (MML) methods, while the parameters of normal ogive models are commonly estimated with Bayesian estimation methods like Markov chain Monte Carlo (MCMC).

In maximum likelihood estimation, item parameters are considered as fixed unknown parameters, while ability can be viewed as either a fixed or a random variable. When ability θ is assumed to be a random latent variable with a continuous distribution $g(\theta)$, marginal maximum likelihood (MML) estimation can be implemented. The unconditional or marginal probability of a response pattern Y_j for a random candidate *j* sampled from a population with ability distribution $g(\theta)$ is obtained as follows

$$P\left(\mathbf{Y}_j\right)=\int_{-\infty}^{+\infty}P\left(\mathbf{Y}_j,\theta\right)d\theta=\int_{-\infty}^{+\infty}P\left(\mathbf{Y}_j\mid\theta\right)g\left(\theta\right)d\theta=\int_{-\infty}^{+\infty}\left[\prod_{i=1}^{I}P_i(\theta_j)^{Y_{ij}}(1-P_i(\theta_j)^{1-Y_{ij}}\right]g\left(\theta\right)d\theta, \quad (1)$$

where Y_{ij} denotes whether the item *i* is answered correctly $(Y_{ij} = 1)$ or not $(Y_{ij} = 0)$. The complete likelihood function can be computed under the assumption of response independence between examinees as

$$P(\mathbf{Y} \mid \theta) = \prod_{j=1}^{J} P(\mathbf{Y}_j \mid \theta). \tag{2}$$

Estimates for item parameters are found by maximizing the log of the likelihood (2), turning out with different likelihood equations depending on the chosen IRT model. In an maximum likelihood framework, uncertainty in the estimated item parameters is represented by the standard errors of the parameter estimators.

A fully Bayesian parameter estimation can be conducted under the assumption that all the quantities of interests (i.e. item parameters and individual abilities) are random variables. The binary response variable Y_{ij} should be treated first by introducing independent and normal identically distributed Y_{ij}^{*}, with expected value equal to $a_i(\theta_j - b_i)$ and unit variance, representing underlying responses and sampled from a normal distribution above 0 when the response is correct or below 0 in case the response is incorrect. The joint posterior distribution of interest is the following

$$P(\mathbf{Y}^{*}, \boldsymbol{\theta}, \boldsymbol{\xi} \mid \mathbf{Y}) = P(\mathbf{Y}^{*} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Y}) P(\boldsymbol{\theta}) P(\boldsymbol{\xi}), \tag{3}$$

where $\boldsymbol{\xi}$ is the vector containing all item parameters and all terms are expressed in matrix notation. The joint posterior distribution in (3) has an intractable form and the Gibbs sampler (Geman & Geman, 1984) can be used in order to draw samples from each single distribution of a variable conditional to all other variables iteratively until convergence. Uncertainty in the item parameters is reflected by the posterior standard deviation. The

95% credibility interval (CI) contains 95% of the highest area of the posterior density, and can be used as a measure of uncertainty of the parameter estimates. CI's can be calculated for item parameters, but also for the amount of information an item provides at a certain ability level, when a Markov chain for the information is built in the Monte Carlo iterations.

Assumptions underlying IRT are often (slightly) violated in practice. Nuisance abilities or fatigue might play a role, which violates the assumption of uni-dimensionality. The assumption of local independence could be violated due to the use of common stimuli for groups of items. Besides, item parameters are always estimated based on finite samples, which results in estimation errors. Because of this, uncertainty, either denoted as standard deviation or as credibility interval, is involved in the estimation process. The question remains, how much this uncertainty affects the information provided by the items.

Uncertainty and item information functions

In order to find out how uncertainty affects information, the uncertainty about the parameters could be modeled explicitly in the item information functions. The item information function, or Fisher Information function, is defined as:

$$I_i(\theta) = -E\left\{\frac{\partial^2 P(Y \mid \theta)}{\partial \theta^2}\right\}. \tag{4}$$

For the 3-parameter logistic model (3PLM) , this expression simplifies to

$$I_i(\theta) = \frac{a_i^{\,2}(1-c_i)}{[c_i + e^{a_i(\theta - b_i)}][1 + e^{-a_i(\theta - b_i)}]^2}. \tag{5}$$

where (a_i, b_i, c_i) represent the discrimination, difficulty and guessing parameter of an item. When uncertainty in the parameters (a_i, b_i, c_i) is modeled explicitly by denoting the parameter values as $(\hat{a}_i + \Delta\hat{a}_i, \hat{b}_i + \Delta\hat{b}_i, \hat{c}_i + \Delta\hat{c}_i)$, where the hat emphasizes that the parameters are estimated, and Δ denotes the uncertainty in each parameter, the item information function could be formulated as

$$I_{i\Delta}(\theta) = \frac{(\hat{a}_i + \Delta\hat{a}_i)^2 (1 - (\hat{c}_i + \Delta\hat{c}_i))}{[(\hat{c}_i + \Delta\hat{c}_i) + e^{(\hat{a}_i + \Delta\hat{a}_i)(\theta - (\hat{b}_i + \Delta\hat{b}_i))}][1 + e^{-(\hat{a}_i + \Delta\hat{a}_i)(\theta - (\hat{b}_i + \Delta\hat{b}_i))}]^2}. \quad (6)$$

To illustrate the effect of uncertainty, the item information function for a hypothetical item *i* with estimated parameters $(\hat{a}_i = 1.4, \hat{b}_i = 0, \hat{c}_i = 0.2)$ is shown in Figure 1a. Let the uncertainty in the parameters be described by the standard error of measurement $(s.e.(\hat{a}_i) = 0.05, s.e.(\hat{b}_i) = 0.1, s.e.(\hat{c}_i) = 0.03)$. Figure 1b shows the item information functions resulting from five random draws, where the item parameters were assumed to follow a normal distribution with means $(\hat{a}_i, \hat{b}_i, \hat{c}_i)$ and standard deviations $(s.e.(\hat{a}_i), s.e.(\hat{b}_i), s.e.(\hat{c}_i))$. Each of these information functions, could have been the original information functions that appeared as the curve in Figure 1a after estimating the item parameters with uncertainties.

==========

Insert Figure 1 at about here

==========

The effects of uncertainty in the item parameters can also be investigated individually. To find the effects of uncertainty in the discrimination parameter $\Delta\hat{a}_i$ on the information function, it has to be noted first that item *i* is most informative for θ in the proximity of $(\hat{b}_i + \Delta\hat{b}_i)$. For these θ-values, it can be derived that the effect of $\Delta\hat{a}_i$ on the denominator is negligible. It can be shown that

$$I_{i\Delta}(\theta) \approx \left(1 + 2\frac{\Delta\hat{a}_i}{\hat{a}_i^2} + \left(\frac{\Delta\hat{a}_i}{\hat{a}_i}\right)^2\right) I_i(\theta) \qquad (7)$$

for θ close to $(\hat{b}_i + \Delta\hat{b}_i)$, and small deviations $\Delta\hat{c}_i$. In other words, uncertainty in the discrimination parameter $\Delta\hat{a}_i$ affects the amount of information provided by the item. When $\Delta\hat{a}_i$ is positive, the discrimination parameter a_i is underestimated, and the amount of information is underestimated as well. On the other hand, when $\Delta\hat{a}_i$ is negative, the discrimination parameter is overestimated, and the amount of information provided by the item is also overestimated.

Uncertainty in the difficulty parameter $\Delta\hat{b}_i$ only plays a role in the denominator of Equation 6. It mainly defines where the item is most informative. For positive $\Delta\hat{b}_i$, the difficulty of the item is underestimated and the item will be more informative for ability levels slightly higher than the difficulty level. For negative $\Delta\hat{b}_i$, the difficulty will be overestimated and the opposite will hold.

The guessing parameter c_i in Equation 5 is related to the amount of information provided by the item. The higher the c_i, the lower the amount of information provided by the item. The uncertainty $\Delta\hat{c}_i$ also affects the amount of information. When the deviation is

positive, the guessing parameter $\hat{c}_i$ is underestimated, and the amount of information provided by the item is overestimated. For negative values of $\Delta\hat{c}_i$, the opposite holds. Finally, to gain insight in the consequences of uncertainty in the item parameters, 95% reliability intervals were drawn around the information function of item *i* for small, medium and large uncertainties in Figure 2.

==========

Insert Figure 2 at about here

==========

Even for small uncertainties in the discrimination and difficulty parameters, the uncertainty in the information function will be substantial. The question remains how to take it into account during the test assembly process.

Impact on automated test assembly and ability estimation.

Hambleton & Jones (1994) studied the impact of uncertainty in the information function in automated test assembly. They took the effects of capitalization on chance during test assembly into account. When optimal test assembly techniques are applied, items with higher levels of discriminating power tend to be selected due to favorable measurement properties (e.g. van der Linden, 2005). These items also have higher chances of positive errors in their item parameter estimates. Due to this selection effect, optimal test assembly might "capitalize on chance" by selecting items with positive estimation errors. When the assembled test is administered and calibrated, the test will be less informative than expected, because of regression towards the mean of the parameter estimates. In their empirical study, Hambleton & Jones found that for small calibration samples,

magnitude of overestimation of the information function ranged up to 104%. They also found that two factors were influencing this capitalization of chance: size of the calibration sample and ratio item bank size to test length. Large calibration samples reduce uncertainty in the parameter estimates, which reduces overestimation of information in the test. Besides, when the ratio item bank size to test length is small, relatively many items are selected from the bank, and item selection does not capitalize that much on items with positive estimation errors.

In this paper, we focus on uncertainty and automated test assembly. To prevent the problems described by Hambleton and Jones (1994), we come up with a different solution. A robust test assembly algorithm is applied that corrects for item parameter uncertainty during test assembly.

Robust optimization

The ultimate goal of robust optimization (Ben Tal, El Ghaoui, & Nemirovski, 2009) is to take data uncertainty, i.e. uncertainty in the item parameters or in the information function, into account already at the item selection stage to "immunize" resulting tests against this uncertainty. Under this approach, we are willing to accept a suboptimal solution for the problem in order to ensure that the solution remains near optimal when the estimated parameters turn out to differ from their real values. When robust optimization is applied to automated test assembly, this implies that we are willing to accept a test that is suboptimal with respect to measurement precision, but that is guaranteed to be near optimal even when the information of the items is overestimated.

To apply robust optimization, an uncertainty set Z hast to be defined that describes the uncertainty in the parameters. A robust model for the construction of a linear test would be:

$$\max y \tag{8}$$

subject to

$$\sum_{i \in I} I_i(\theta, \zeta) x_i \geq y \quad \forall \zeta \in \mathrm{Z}, \tag{9}$$

$$\sum_{i \in S_c} x_i \leq b_c \quad \forall c, \tag{10}$$

$$\sum_{i=1}^{I} q_j x_i \leq b_q \quad \forall q, \tag{11}$$

$$\sum_{i \in e} x_i \leq 1 \quad \forall e, \tag{12}$$

$$\sum_{i=1}^{I} x_i = n, \tag{13}$$

$$x_i \in \{0,1\}. \tag{14}$$

where x_i indicates whether the item is selected (x_i =1) or not (x_i =0), $I_i(\theta,\zeta)$ denotes the information function of item i at ability level θ, ζ denotes the uncertainty, and Z is the uncertainty set. Equations 10 till 12 denote the categorical, quantitative, and enemy constraints (van der Linden, 2005, chap. 3.2). The test length is set equal to n items in Equation 13. Finally Equation 14 defines the decision variables. The difference between linear test assembly and robust test assembly is in Equation 9, where the uncertainty set has been added to the formulation.

Different algorithms for solving the problem in Equations 8 - 14 have been proposed. First of all, Soyster (1973) developed an algorithm, where all parameters in the uncertainty set were fixed at their minimal values. In Soyster's algorithm the test assembly model in Equations 8 – 14 slightly changes. Equation 9 is reformulated as:

$$\sum_{i \in I} \inf_{\zeta \in Z} I_i(\theta, \zeta) x_i \geq y. \qquad (9')$$

Soyster's approach is very conservative, because the probability that all parameters take the minimum value of their CI's is extremely small. Especially for large problems, over-conservatism reduces the practical value of this method. In order not to be too conservative, we follow De Jong, Steenkamp, & Veldkamp (2009) and propose a modified version of Soyster's algorithm for dealing with robust optimization problems, by fixing the psychometric parameters at the estimated values minus one (posterior) standard deviation, instead of fixing the parameters at their infima. Equation 9 in the test assembly model now changes into:

$$\sum_{i \in I} \overline{I_i(\theta)} x_i \geq y \qquad (9'')$$

where $\overline{I_i(\theta)}$ equals the estimated information function at ability level θ minus one (posterior) standard deviation. A robust optimization algorithm for automated test assembly, can be described as follows:

1. Obtain an estimate of the item parameters and the uncertainty in them.
2. Formulate the 0-1 LP model for automated test assembly
3. Specify a relevant grid of theta values across the ability range of interest, denote them by θ_m.

4. Calculate the robust approximation of the information for each of these gridpoints.
5. Insert the robust approximation of the information function in the model formulated in (2).
6. Solve the test assembly problem modeled in (2).

Robust approximations of the information function in step 4 of the algorithm can be calculated by applying multiple imputation (Rubin, 1987). The item parameters play the role of missing data, but they have a multivariate normal distribution where the item parameter estimates are the mean and the uncertainty defines the variance. Draw *K* item parameter vectors $\xi_k = (a_k, b_k, c_k)$, and calculate the information $I(\xi_k, \theta_m)$ for each of the *m* gridpoints defined in step 3. Finding the robust approximation that subtracts one standard deviation from the information function is equivalent to finding the 15.78^{th} percentile of the distribution of $I(\xi_k, \theta_m)$ for each gridpoint *m*. Within a Bayesian framework, the robust approximation might also be derived from the Markov chains generated during the parameter estimation process.

To solve both the standard test assembly problem and its robust counterpart formulated in Step 5 of the algorithm, several standard methods for solving 0-1 LP problems can be applied (van der Linden, 2005, Chap. 4). In the empirical examples, we used a Simulated Annealing heuristic (see also, Veldkamp, 2002), but other methods can be applied.

Empirical examples

The impact of the modified version of Soyster's algorithm on test assembly was studied in the first example both for uncertainty in the discrimination parameters (Example 1a) and for uncertainty in the difficulty parameters (Example 1b). The implementation of the robust test assembly algorithm was illustrated in the second example, where the different steps of the algorithm were followed, to assemble a robust test with maximum information around a cut-off point. In both examples, application of the modified version of Soyster's algorithm was compared with the application of 0-1 linear programming test assembly. In the first example, items were used with a large uncertainty in the item parameters, where in the second example items were used with a small uncertainty in the parameters.

Example 1: Uncertainty in the item parameters.

The Connector Ability (Maij- de Meij, Schakel, Smid, Verstappen, and Jaganjac, 2008) is a test battery for measuring Intelligence that consists of three different subtests: Number series, Figure series, and Raven's matrices. The modified version of Soyster's algorithm was applied to assemble a 20-item linear Number Series test.

For the Number series subtest, an item bank of 253 items was available. The MIRT software package, developed at the University of Twente in the Netherlands, was applied to calibrate the items with the 2PL model based on a sample of 3000 respondents, where each of the respondents answered 20 items in a balanced block design. The resulting parameter values and their standard errors (SE) are shown in Figure 3a and Figure 3b.

==========

Insert Figure 3a and Figure 3b at around here.

==========

The discrimination parameters (*a*) are in the interval [0.25, 2.29] with a mean value of $\mu_a = 1.14$ and a mean value of the standard error $\mu_{SE(a)} = 0.21$. The correlation between the discrimination parameters and the SE's was equal to $\rho_{a,SEa} = 0.78$. This implies that the discrimination parameters of the more discriminating items have been estimated with larger errors. The difficulty parameters (*b*) are in the interval [-2.33, 1.69], with a mean value of $\mu_b = -0.44$ and a mean value of the standard error $\mu_{SE(b)} = 0.20$. Figure 4b shows a quadratic relation between the difficulty parameter and the SE's. This implies that the difficulty parameters in the middle of the distribution were estimated with the highest precision.

The relation between the difficulty parameters *b* and the standard errors of the discrimination parameter is shown in Figure 4. It can be seen that the discrimination parameters of the easy items have been estimated with the highest standard errors.

==========

Insert Figure 4 at around here.

==========

Uncertainty in the discrimination parameter

To illustrate the consequences of taking the uncertainty in the discrimination parameters *a* into account, four 20-items tests were assembled from the NS item bank. Each test consisted of the 20 most informative items at four different ability levels ($\theta \in \{-2,-1,0,1\}$). It should be noted that the Connector Ability is an adaptive test, but for the purpose of illustrating robust optimization techniques, four linear tests were assembled. First, the

tests were assembled without taking the uncertainty into account. Then a modified version of Soyster's algorithm was applied, where the one standard error was subtracted from each discrimination parameter *a*. The resulting test information functions are shown in Figure 5.

==========

Insert Figure 5 at around here.

==========

For low ability respondents (θ = -2), the loss of information is substantive (37%) when the uncertainty in *a* is taken into account. For (θ = -1) and (θ = 0) the loss is 25%, and for high ability respondents (θ = 1), the loss is 20%. For all four cases it can be concluded that the measurement precision of the test might be overestimated when the uncertainty in the parameters is not taken into account. The reason why the low ability respondents lose most information is that the discrimination parameters of the easy items have been estimated with highest uncertainty (see Figure 4).

Item overlap between the four tests and their robust counter parts varied from 80% (θ = -2) to 95% (θ = 1), which implies that almost the same set of items is selected in 0-1 linear programming and Robust test assembly. Further analyses revealed that the order in which the items were selected was quite different. Especially for low theta values (θ = -2) and for high theta values (θ = 1) the change in position was on average 3 or 4 positions when the items were ordered with respect to the amount of information they provide.

Uncertainty in the difficulty parameter

Application of the modified version of Soyster's algorithm to take uncertainty in the difficulty parameter into account, would imply that

$$\begin{aligned}\overline{I(\theta)} &= \min_{\zeta \in Z} I(\theta, \zeta) \\ &= \min_{\bar{b} \in [b - sd_b, b + sd_b]} I(\theta; a, \bar{b}) \\ &= \begin{cases} I(\theta; a, b + sd_b) & \theta \le b, \\ I(\theta; a, b - sd_b) & \theta > b. \end{cases}\end{aligned} \tag{15}$$

In other words, a robust information function that takes uncertainty in the difficulty parameter into account can be derived from the information functions based on the lowest and highest difficulty values in the uncertainty interval. Again, four 20-items tests were assembled from the NS item bank. Each test consisted of the 20 most informative items at four different ability levels ($\theta \in \{-2,-1,0,1\}$). The results are shown in Figure 6.

==========

Insert Figure 6 at around here.

==========

It can be seen that the differences in test information function between robust test assembly and its 0-1 LP counterpart are relatively small. The overlap in selected items was 95% or higher, and even the order in which the items were selected was almost the same for both test assembly algorithms.

Example 2: Uncertainty in the information function.

For the second example, an item bank for the Logical Reasoning (LR) section of the Law School Admission Test (LSAC, 2010) was used. 150 items were calibrated with a 3PLM using BILOG MG 3, for a sample of 41,500 candidates. The estimated item parameters

ranged from $\hat{a}_i \in [0.22,1.22]$, $\hat{b}_i \in [-3.14,2.24]$, and $\hat{c}_i \in [0.0,0.49]$, and the average uncertainty $(\overline{\Delta\hat{a}}, \overline{\Delta\hat{b}}, \overline{\Delta\hat{c}})$ was equal to $(.020,.043,.014)$. These uncertainties are relatively small, since the item parameters were estimated based on a large sample of examinees. Nine different item types were distinguished. The goal was to assemble a 20-item test that was maximum informative around $\theta = 0,$ and where a number a predefined number of items was selected for each of the item types. The problem could be modeled as:

$$\max \sum_{i=1}^{150} I_i(\theta = 0) x_i \tag{16}$$

subject to

$$\sum_{i \in V_j} x_i \leq b_j \quad j = 1,..,9, \tag{17}$$

$$\sum_{i=1}^{150} x_i = 20, \tag{18}$$

$$x_i \in \{0,1\} \quad i = 1,..,150, \tag{19}$$

where V_j denotes the subset of items with item type *j*, and b_j denotes the maximum number of items to be selected for this category. It has to be mentioned that these specifications do not resemble the specifications for the LR section of the LSAT.

For this specific test assembly problem, there was only one ability point of interest ($\theta = 0$). Robust values for the information at ($\theta = 0$) were obtained by generating 500 random draws from a normal distribution with means $(\hat{a}_i, \hat{b}_i, \hat{c}_i)$ and standard deviations $(s.e.(\hat{a}_i), s.e.(\hat{b}_i), s.e.(\hat{c}_i))$, for each of the 150 items in the pool. For each draw the information at ($\theta = 0$) was calculated. The resulting values were ranked for each of the items and the values of the $15,78^{th}$ percentile were selected as robust approximations that

could be used in the modified version of Soyster's approach. The robust approximations of the information function were implemented in the test assembly model. Both the 0-1 LP test assembly problem in (16) – (19) and its robust counterpart were solved. Resulting tests are shown in Figure 7.

==========

Insert Figure 7 at around here.

==========

It can be seen in this example that application of the modified Soyster's approach results in a decrease of information by 5%, even for an item bank where uncertainty in the item parameters is very small. For both the 0-1 LP and the modified version of Soyster's approach the same set of items was selected for the test.

Discussion

Even though Hambleton & Jones (1994) already demonstrated how uncertainty in the item parameters might result in overestimation of the information in test, until now, these uncertainties are not reckoned with in operational test assembly algorithms. Instead, it is often argued that the uncertainties in the parameters are normally distributed and that they will cancel out over all items in the test. This argument might sound appealing, but it does not take into account that optimal item selection capitalizes on positive estimation errors, which results in overestimation of the amount of information in the tests. To emphasize the impact of the problem, it was described into detail how uncertainty enters test assembly problems due to estimation errors in various estimation procedures. When

the effects of uncertainty were examined, both analytically and graphically, it was demonstrated that even small uncertainties would have quite an impact on the item information functions.

In the numerical examples, it was demonstrated that 0-1 LP was most sensitive to uncertainties in the discrimination parameters. These examples also illustrated that not taking the uncertainties into account might result in an overestimation of the amount of information in the test, which implies a serious threat for the validity of the resulting scores. For large uncertainties, the decrease of information in robust test assembly was up to 37%, whereas for small uncertainties, the loss of information was still around 5%.

It might sound strange to claim that robust tests with less information are actually more reliable than 0-1 LP based tests. But in this paper, it was demonstrated how uncertainty in the item parameters accounts for this effect. Not taking the uncertainty into account results in overestimation of the information and as a result, in overestimation of the reliability of the test. In case bounds have been imposed on the amount of information in the test, uncertainty might also result in violation of these bounds, and feasibility problems might arise during test assembly (Huitzing, Veldkamp, & Verschoor, 2007). Operational item banks are often calibrated based on relatively small samples and, like Hambleton and Jones (1994) already demonstrated, items calibrated with small samples are most vulnerable to overestimation of the information. In general, binary IRT models can be estimated with a reasonable fit, based on a sample of 1000 candidates. However, the first example illustrates that the uncertainty resulting from such small samples might have a considerable impact on test assembly. The second example demonstrates that even for large samples, uncertainty cannot be neglected.

For practical purposes, it has to be remarked that the mode of test administration influences the impact of uncertainty, as well. In CAT only one item is administered at a time, and capitalization of chance is a serious problem. In linear testing, it is less of a problem, since the ratio item bank – test length is much smaller. But we also recommended to use the modified version of Soyster's algorithm for the assembly of linear tests, when test length is smaller than 15,78% of the item bank size, since at least 15,78% of the items will have an uncertainty in the information function larger than the amount of uncertainty that was taken into account.

Implementation of the modified version of Soyster's algorithm turned out to be a rather straightforward way to deal with uncertainties in the item parameters. Unfortunately, the method assumes the same amount of uncertainty in all the items, which is not very realistic. Uncertainty in some items will be much higher, while uncertainty in most items will be lower. Subtracting one standard deviation for all items is just a compromise, but for long linear tests it might be too conservative. It is a topic of further research to find out whether applications of more recent approaches to robust optimization, like the methods proposed by Ben-Tal and Nemirovski (2000) or by Bertsimas and Sim (2004), would reduce the gap between robust optimization and 0-1 LP or that it would underline the conclusion that application of 0-1 LP result in a serious overestimation of the amount of information in the test.

Even though uncertainty in item parameters is not taken into account directly, some automated test assembly methods do pay attention to it, especially in the area of CAT. The alpha-stratification method (Chang & Ying, 1999) for CAT was originally developed to prevent overexposure of items with highest discrimination parameters. But indirectly,

this method also prevents overexposure of items with highest probability of positive estimation errors for the discrimination parameters, and it reduces uncertainty in the outcomes of the test. Besides, in the Weighted Deviation Method (WDM) (Stocking & Swanson, 1993), specifications are formulated as weak constraints that might be violated during test assembly. One of the reasons to impose weak constraints is that it is not very realistic to impose strict constraints when some parameters are estimated with uncertainty. The comparison of robust test assembly with these methods for CAT might be an interesting topic for further research.

Finally, it should be remarked that dealing with uncertainties due to parameter estimation is only one of the virtues of robust optimization. Recently, several methods for automated item generation were proposed. Matteucci, Mignani & Veldkamp (2012) presented an approach based on regression trees and Glas & van der Linden (2003) and Glas, van der Linden, & Geerlings (2010) proposed an item cloning technique for generating new items. Both methods predict the parameters of the generated items. The predictions parameters can be used to start a computerized adaptive test and the parameters can be estimated on the fly. In this way, the time-consuming and expensive calibration of the item pool can be shortened or even skipped. Robust test assembly methods that take the uncertainties due to parameter prediction instead of estimation into account, can contribute to the validity of such an approach.

References

Armstrong, R.D., Jones, D.H., & Wang, Z. (1995). Network optimization in constrained standardized test construction. *Applications of Management Science, 8,* 189-212).

Barrada, J.R., Abad, F.J., & Veldkamp, B.P., (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema, 21,* 313-320.

Barrada, J.R., Olea, J., & Ponsoda, V. (2007). Methods for restricting maximum exposure rate in computerized adaptive testing. *Methodology, 3,* 14-23.
DOI 10.1027/1614-1881.3.1.14

Barrada, J.R., Veldkamp, B.P., & Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement, 33,* 58-73.
DOI: 10.1177/0146621608315329

Belov, D.I., & Armstrong, D.H. (2005). Monte Carlo test assembly for item pool analysis and extension. *Applied Psychological Measurement, 29,* 239-261.
DOI: 10.1177/0146621605275413

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust Optimization.* Princeton, NJ: Princeton University Press.

Ben-Tal, A., Nemirovski, A. (2000). Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming, 88,* 175-184. DOI: 10.1007/s101070000163

Bertsimas, D., Sim, M. (2004). The price of robustness. *Operations Research, 52,* 35-53. DOI: 10.1287/opre.1030.0065

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Chang, H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23,* 211-222. DOI: 10.1177/01466219922031338

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

De Jong, M.G., Steenkamp, J.-B.G.M., & Veldkamp, B.P. (2009). A Model for the Construction of Country-Specific Yet Internationally Comparable Short-Form Marketing Scales. *Marketing Science, 28,* 674-689. DOI:10.1287/mksc.1080.0439

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721-741.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27,* 247-261.
DOI: 10.1177/0146621603254291

Glas, C. A. W., & van der Linden, W. J., & Geerlings, H. (2010). Estimation of the parameters in an item-cloning model for adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 289-314). New York: Springer.

Hambleton, R.H., & Jones, R.W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 7,* 171-186. DOI: 10.1207/s15324818ame0703_1

Huitzing, H.A., Veldkamp, B.P., & Verschoor, A.J. (2005). Infeasibility in Automated Test Assembly Models: A Comparison Study of Different Methods. *Journal of Educational Measurement, 42,* 223-243. DOI: 10.1111/j.1745-3984.2005.00012.x

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph, 7.*

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

LSAC (2010). *The official LSAT Handbook.* Newtown, PA: Law School Admission Council, Inc.

Luecht, R.M., & Hirsch, T.M. (1992). Item selection using an average growth approximation of target information functions. *Applied Psychological Measurement, 16*, 41-51. DOI: 10.1177/014662169201600104

Matteucci, M, Mignani, S., & Veldkamp, B.P. (2012). Prior Distributions for Item Parameters in IRT Models. *Communication in Statistics, Theory and Methods, 41*. in press.

Maij- de Meij, A.M., Schakel, L., Smid, N., Verstappen, N., & Jaganjac, A. (2008). *Connector Ability; Professional Manual*. Utrecht, The Netherlands: PiCompany B.V.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

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

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 311-327. DOI: 10.1111/j.1745-3984.1998.tb00541.x

Rubin, D.B. (1987). *Multiple imputations for nonresponse in surveys.* New York: Wiley.

Soyster, A.L. (1973). Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research, 21,* 1154-1157.

Stocking, M.L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277-292. DOI: 10.1177/014662169301700308

Sympson, J.B., & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

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

van der Linden, W.J., & Veldkamp, B.P. (2004). Constraining Item Exposure Rates in Computerized Adaptive Testing with Shadow Tests. *Journal of Educational and Behavioral Statistics, 29,* 273-291. DOI: 10.3102/10769986029003273

van der Linden, W.J., & Veldkamp, B.P. (2007). Conditional item exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32,* 398-417. DOI: 10.3102/1076998606298044

Veldkamp, B.P. (2002). Multidimensional Constrained Test Assembly. *Applied Psychological Measurement, 26,* 133-146. DOI: 10.1177/01421602026002002

Verschoor, A.J. (2007). *Genetic algorithms for automated test assembly.* Unpublished doctoral dissertation. Enschede, The Netherlands: University of Twente.

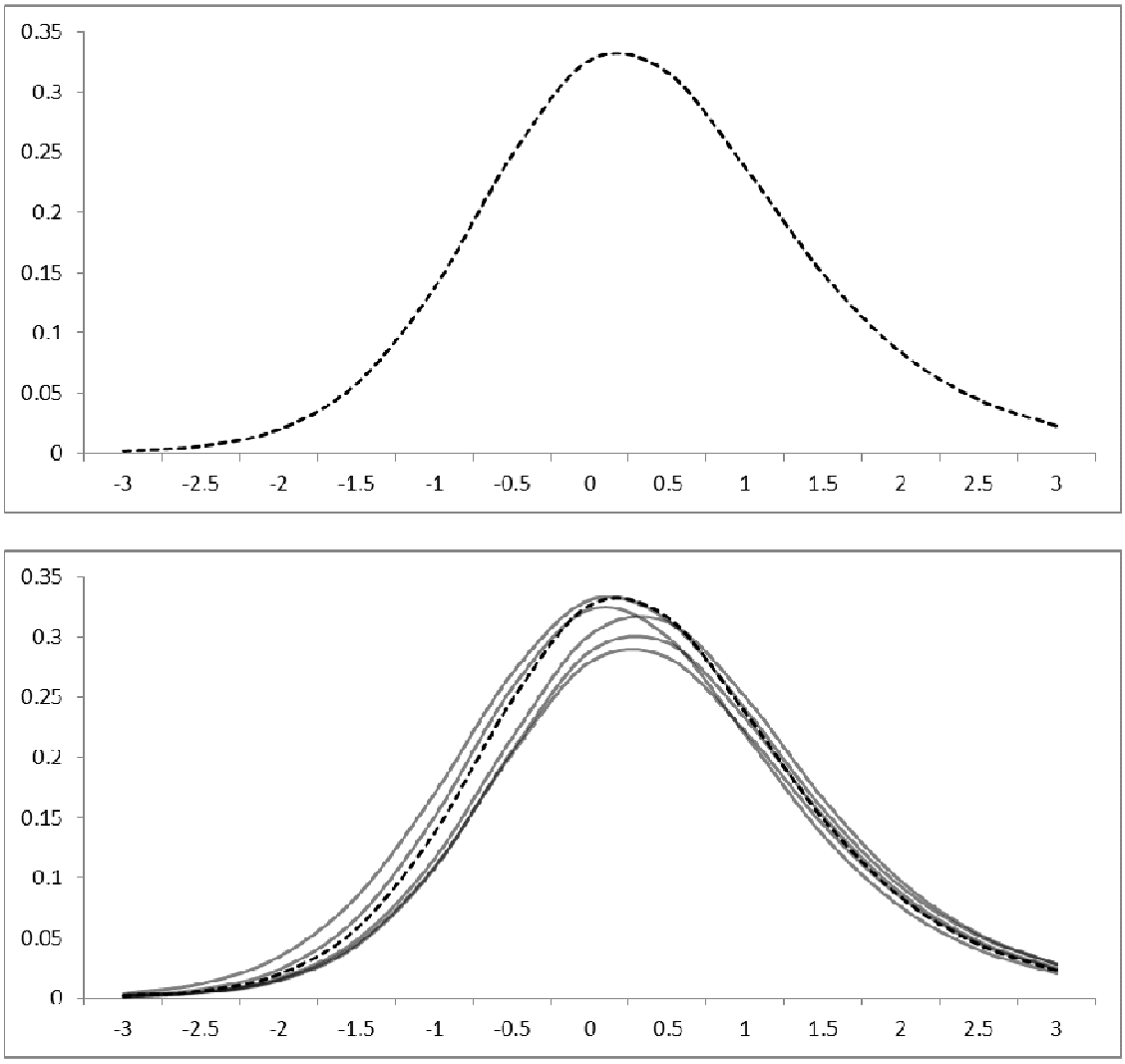

Figure 1a: Item information function $(a_i = 1.4, b_i = 0, c_i = 0.2)$. *Figure 1b:* Item information functions from the same distribution.

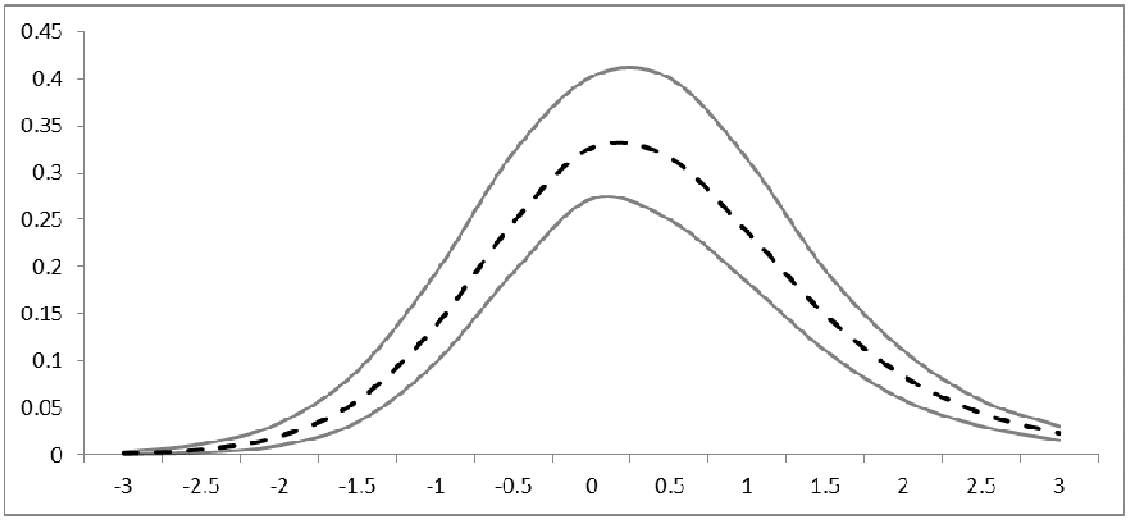

Figure 2: 95% Reliability Interval for the item information function

$(a_i = 1.4, b_i = 0, c_i = 0.2)$.

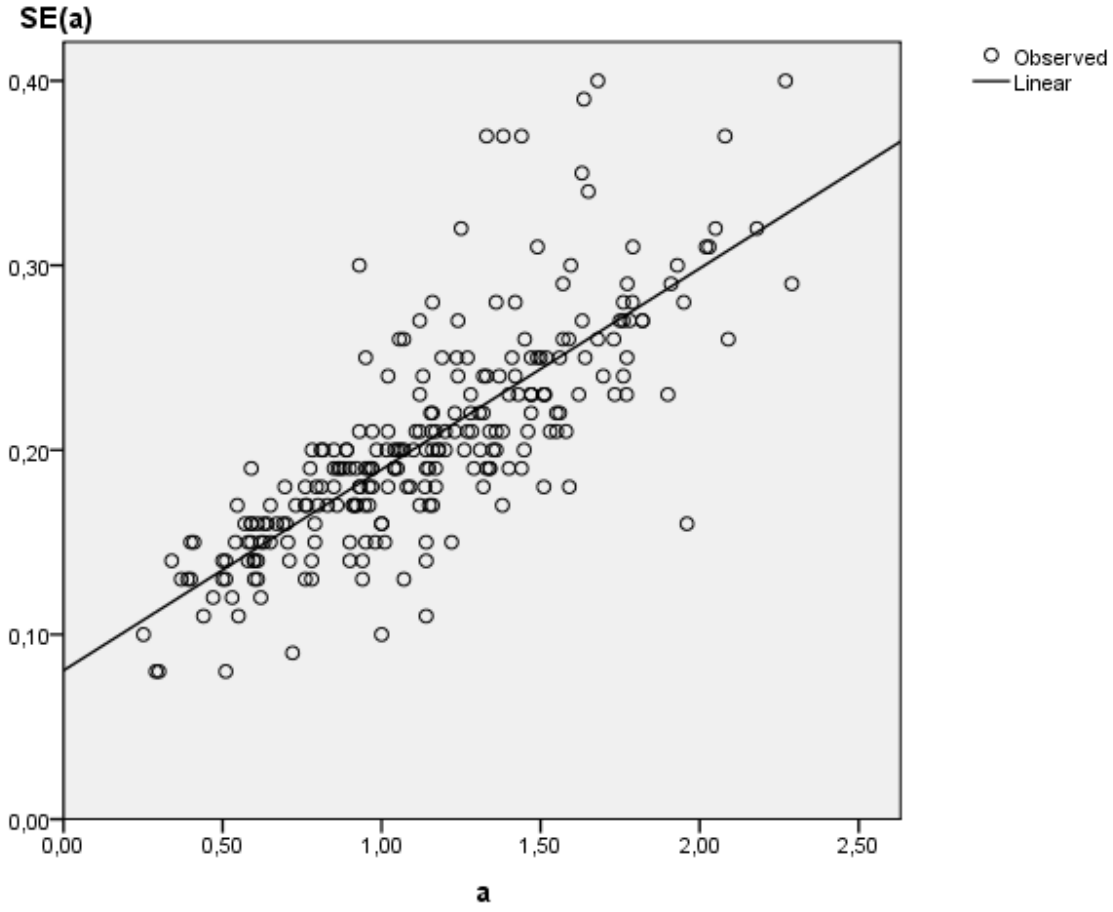

Figure 3a: Relation between discrimination parameter *a* and its standard error

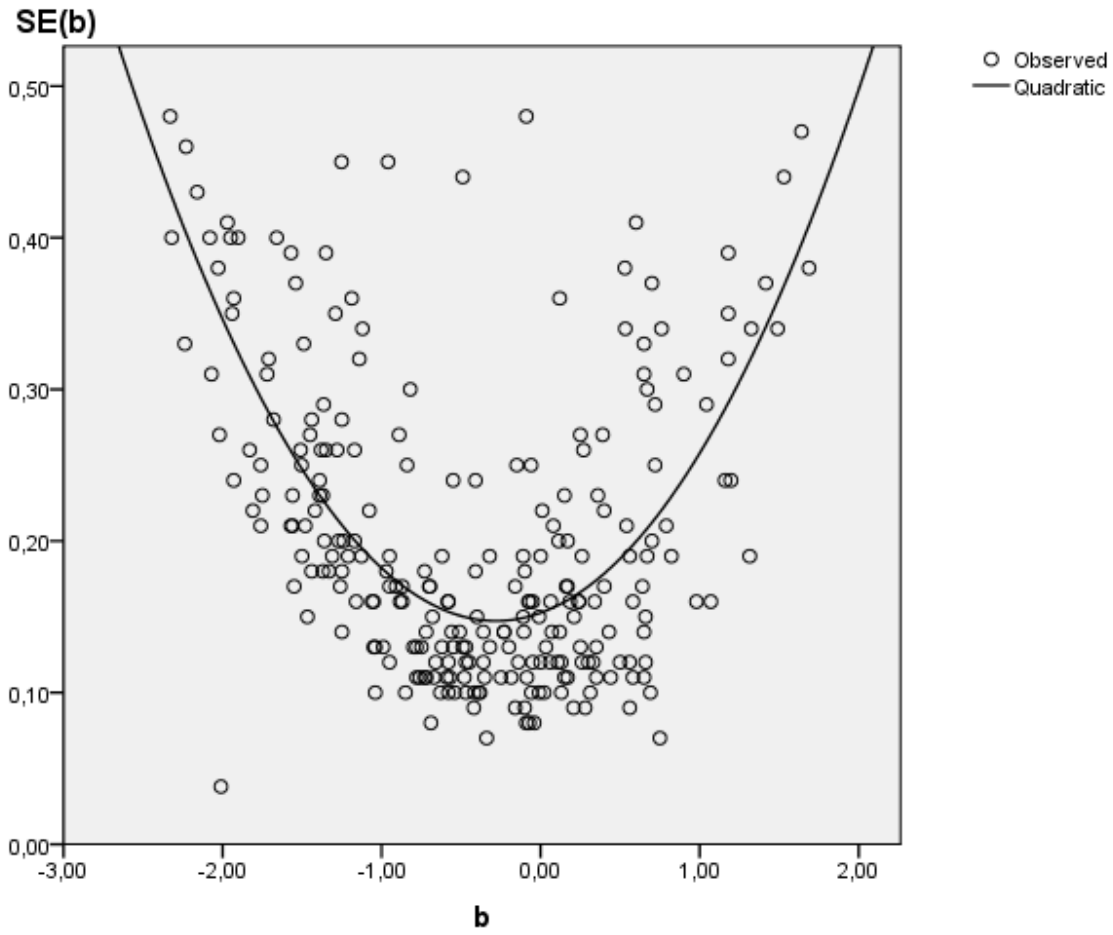

Figure 3b: Relation between difficulty parameter *b* and its standard error

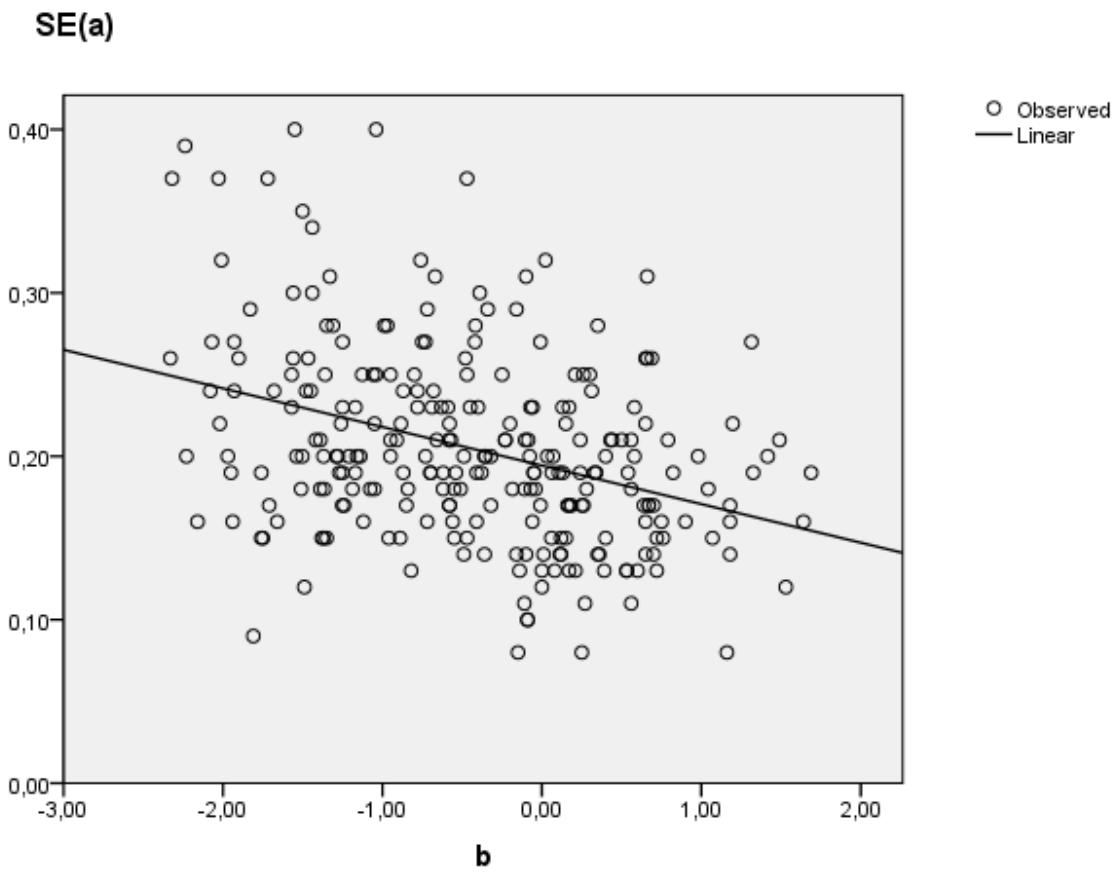

Figure 4: Relation between difficulty parameter *b* and the standard error of the discrimination parameter *a*.

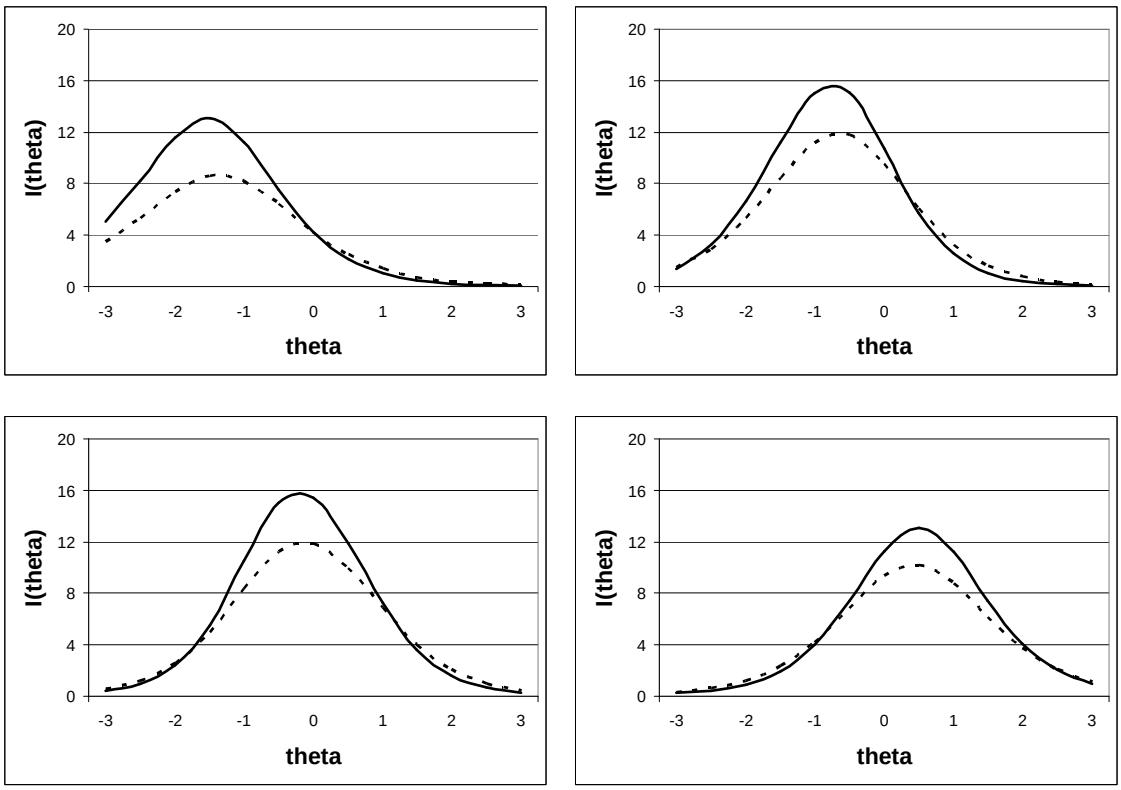

Figure 5: Test information functions without (solid line) and with (dotted line) taking the uncertainty in the discrimination parameter into account

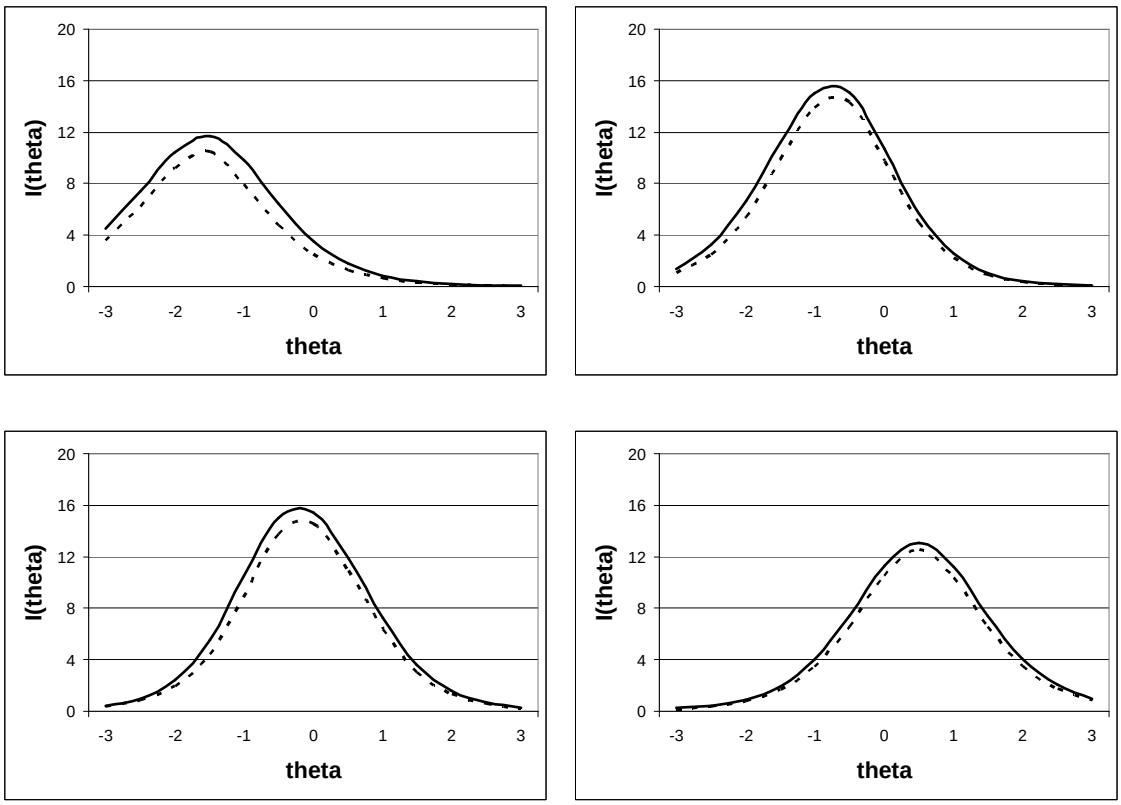

Figure 6: Test information functions without (solid line) and with (dotted line) taking the uncertainty in the difficulty parameter into account

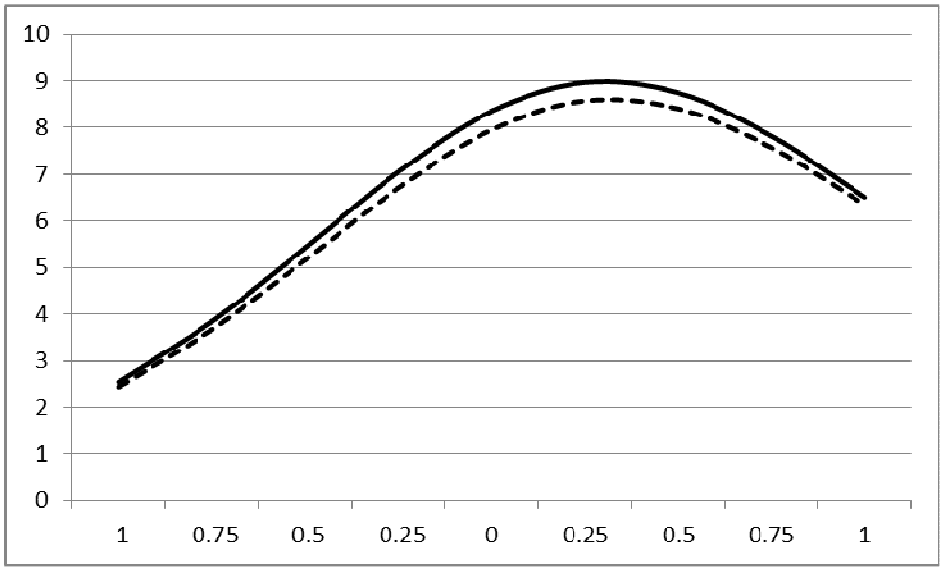

Figure 7: Test information function without (solid) and with (dashed) taking uncertainty in the item parameters into account when uncertainty in the parameters is small.