Chapter 15
Robust Computerized Adaptive Testing

Bernard P. Veldkamp and Angela J. Verschoor

Abstract In order to optimize measurement precision in computerized adaptive testing (CAT), items are often selected based on the amount of information they provide about a candidate. The amount of information is calculated using item- and person parameters that have been estimated. Usually, uncertainty in these estimates is not taken into account in the item selection process. Maximizing Fisher information, for example, tends to favor items with positive estimation errors in the discrimination parameter and negative estimation errors in the guessing parameter. This is also referred to as capitalization on chance in adaptive testing. Not taking the uncertainty into account might be a serious threat to both the validity and viability of computerized adaptive testing. Previous research on linear test forms showed quite an effect on the precision of the resulting ability estimates. In this chapter, robust test assembly is presented as an alternative method that accounts for uncertainty in the item parameters in CAT assembly. In a simulation study, the effects of robust test assembly are shown. The impact turned out to be smaller than expected. Some theoretical considerations are shared. Finally, the implications are discussed.

15.1 Introduction

In computerized adaptive testing (CAT), the items are administered such that the difficulty level is tailored to the test taker's ability level. Adaptive testing turns out to entail a number of advantages. Candidates only have to answer items that are paired to their ability level, test length can be reduced in comparison to linear test forms, and test administration can be more flexible in terms of time and location as a result of individualized testing. CATs could be offered continuously, on flexible locations, on portable devices, and even via the Web. The advantages of CAT are very appealing for

B. P. Veldkamp (✉)
University of Twente, Enschede, The Netherlands
e-mail: b.p.veldkamp@utwente.nl

A. J. Verschoor
Cito, Arnhem, The Netherlands

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_15

candidates who live in a 21st century world, with tablets, mobile phones and who are continuously online. Computerized adaptive testing is a more and more popular test administration mode in educational, psychological, and health measurement. Many algorithms for tailoring the difficulty level of the test to the individual's ability level have been proposed in the literature (e.g. Eggen 2004, page 6). These algorithms generally consist of the following steps:

1. Before testing begins, the ability estimate of the candidate is initialized (e.g., at the mode of the ability distribution, or based on historical data).
2. Items are selected from an item bank to be maximally informative at the current ability estimate. Sometimes, a number of specifications related to test content or other attributes have to be met, which restricts the number of items available for selection. In this step, an exposure control method is commonly applied to prevent overexposure of the most popular items.
3. Once an item is selected, it is administered to the candidate.
4. The responses are scored.
5. An update of the ability estimate is made after each administration of an item.
6. Finally, the test ends whenever a stopping criterion has been met, for example when a fixed number of items have been administered or when a minimum level of measurement precision has been obtained.

One of the prerequisites of CAT is that a calibrated item pool is available and that the item parameters have been estimated with enough precision to be treated as fixed values. These parameters are used during test administration to calculate the amount of information each item provides and to estimate the ability levels. Unfortunately, item parameters are calibrated with a finite sample of candidates. The resulting item parameter estimates might be unbiased, but they still contain measurement error. This measurement error, which causes uncertainty in the true values of the item parameters, is a source of concern. Previous research on item calibration error in adaptive testing (van der Linden and Glas 2000) already mentioned that items with high discrimination parameters tend to be selected more often from the bank, when items are selected based on the amount of information they provide at the estimated ability level. Especially, positive estimation errors in the discrimination parameters have quite some impact on the amount of information provided. Overestimation of item discrimination will increase the probability that the item will be selected. This phenomenon is also referred to as the problem of capitalization on chance.

Both Tsutakawa and Johnson (1990) and Hambleton and Jones (1994) already studied the effects of item parameter uncertainty on automated assembly of linear test forms. Hambleton and Jones found out that not taking the uncertainty into account resulted in serious overestimation (up to 40%) of the amount of information in the test. To illustrate the effect, Veldkamp (2013) illustrated this with a simulated item pool that consisted of 100 items. The parameters for all of these items were drawn from the same multivariate item parameter distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}^2)$. The mean values of $\boldsymbol{\mu}$ were equal to the true item parameters of a parent item. The discrimination parameter of the parent was equal to $a = 1.4$, the difficulty parameter equal to $b = 0.0$, and the guessing parameter equal to $c = 0.2$. The variance covariance matrix $\boldsymbol{\Sigma}$ was equal

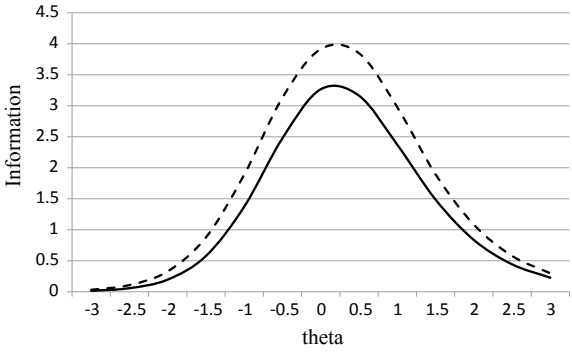

Fig. 15.1 Test information function: ATA (dashed line) or true (solid line)

to the diagonal matrix with the standard errors of estimation ($SE\,a = 0.05$, $SE\,b = 0.10$, $SE\,c = 0.02$) on the diagonal. Because of this, the item parameters only varied due to uncertainty in the parameter estimates. Resulting parameters fell in the intervals $a \in [1.29, 1.52]$, $b \in [-0.31, 0.29]$, and $c \in [0.14, 0.28]$. To illustrate the effects of item parameter uncertainty, a linear test of ten items was selected from this item bank. Fisher information at $\theta = 0.0$ was maximized during test assembly. A comparison of test information functions was made between this test and a test consisting of 10 items with parameter equal to the parameters of the parent item. The results are shown in Fig. 15.1. As can be seen, the test information is overestimated by 20% when uncertainty due to simulated item calibration errors was not taken into account.

Hambleton and Jones (1994) demonstrated that the impact of item parameter uncertainty on automated construction of linear tests depended on both the calibration sample size and the ratio of item bank size to test length. When their findings are generalized to computerized adaptive testing, the impact of calibration sample size is comparable. Calibration error will be larger for smaller samples. For the ratio of item bank size to test length the effects are even larger. In CAT, only one item is selected at a time. The ratio of item pool size to test length is therefore even less favorable. Van der Linden and Glas (2000), studied the impact of capitalization on chance for various settings of CAT in an extensive simulation study, and they confirmed the observations of Hambleton and Jones (1994). In other words, capitalization on chance is a problem in CAT when items are selected based on the amount of information they provide. As a result, the measurement precision of the CATs might be vastly overestimated. Item selection algorithms, therefore have to be modified to account for capitalization on chance.

15.2 Robust Test Assembly

Automated test assembly problems can be formulated as mixed integer linear programming problems. An extensive introduction on how to formulate the mixed integer linear programming problems can be found in van der Linden (2005). These mixed integer programming problems have a general structure where one feature of the test is optimized and specifications for other features are met. For example, the amount of information can be maximized while specifications with respect to the content, the type of items, and the test length have to be met. When a mixed integer linear programming approach is used, the parameters in the model are assumed to be fixed. Due to, for example, calibration error, there is some uncertainty in the parameters and robust optimization methods have to be used. The general idea underlying robust optimization is to take uncertainty into account when the problem is solved in order to make the final solution immune against this uncertainty (Ben Tal et al. 2009).

One of the early methods to deal with item parameter uncertainty in optimization problems was proposed by Soyster (1973). He proposed a very conservative approach, where each uncertain parameter in the model was replaced by its infimum. In this way, a robust lower bound to the solution of the optimization problem could be found. The resulting lower bound turned out to be very conservative though, since it assumed a maximum error in all the parameters, which is extremely unlikely to happen in practice. A modified version of this method was applied to automated assembly of linear tests by de Jong et al. (2009). They took uncertainty due to calibration error into account. The calibration errors were assumed to be normally distributed. But instead of using the infima of these distributions, they subtracted one posterior standard deviation from the estimated Fisher information as a robust alternative. This approach was even studied more into detail by Veldkamp et al. (2013), who studied the effects of uncertainties in various item parameters on Fisher information in the assembly of linear test forms.

A more realistic method to deal with uncertainty in optimization problems was proposed by Bertsimas and Sim (2003). They noted that it almost never happens in practice that uncertainty plays a role for all of the parameters in the model. Instead, uncertainty in a few of the parameters really affects the final solution. They proposed an optimization method where uncertainty only plays a role for Γ of the parameters. For this situation, they proved that finding an optimal solution when at most Γ parameters are allowed to change, is equal to solving $(\Gamma + 1)$ mixed integer optimization problems. In other words, this robust optimization method will be more time consuming, but we can still apply standard software for solving mixed integer programming methods. In automated test assembly, calibration errors are assumed to be normally distributed, and extreme overestimation or underestimation of the item parameters is only expected for a few items the item pool. This resembles the observations of Bertsimas and Sim that uncertainty only affects the final solution for some of the parameters. Therefore, the mixed integer optimization methods for automated test assembly proposed in van der Linden (2005) can still be applied, although the

test assembly models are more complicated and more time consuming to solve. For an application of Bertsimas and Sim (2003) to linear test assembly, see Veldkamp (2013).

15.3 Robust CAT Assembly

Good results were obtained for some practical test assembly problems with the modified Soyster method (see de Jong et al. 2009) and the Bertsimas and Sim method (see Veldkamp 2013). Both methods replace estimated item parameters by a more conservative value either for all or for some of the items, by subtracting one or three standard deviations. These robust optimization methods originate from the field of combinatorial optimization.

A different approach, that originated in the field of psychometrics, can be found in Lewis (1985) where expected response functions (ERFs) are proposed to correct for uncertainty in the item parameters (Mislevy et al. 1994) in the process of constructing fixed-length linear tests. To apply the approach to CAT assembly, ERFs have to be apply at the item pool level. This might result in a starting point for a robust CAT assembly procedures.

15.3.1 Constructing a Robust Item Pool

The calibration error follows a normal distribution. When the distribution of the errors is used, it can be derived which percentage of items will have a certain deviation from the mean. Straightforward application of the cumulative normal distribution illustrates that for 2.5% of the items, a larger deviation than 1.96 times the standard deviation is expected. When the assumption is being made that uncertainty hits were it hurts most, all the items in the pool can be ordered based on the maximum amount of information they provide for any ability value, and expected deviation is subtracted from the estimated information. This can be formulated as:

$$I_i^R(\theta) = I_i(\theta) - z_i * SD(I_i(\theta)), \quad i = 1, \ldots, I, \tag{15.1}$$

where i is the index of the item in the ordered bank, I is the number of items in the bank, $I_i^R(\theta)$ is the robust information provided at ability level θ, z_i corresponds to the $100 \cdot i/(I+1)$-th percentile of the cumulative normal distribution function, and $SD(I_i(\theta))$ is the standard deviation of the information function based on estimated item parameters. A comparable procedure can be applied in a Bayesian framework, however, to calculate z_i the posterior distribution has to be used.

15.3.2 Numerical Example to Illustrate the Concept of Robust Item Pools

An operational item pool of 306 items can be used to illustrate the effects of expected response function, or in our application, robust response function. The items can be calibrated with a three-parameter logistic model (3PLM):

$$P_i(\theta) = c + (1 - c)\frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}, \tag{15.2}$$

where a is the discrimination, b is the difficulty, and c is the guessing parameter. BILOG MG 3 was applied to estimate all item parameters, based on a sample of 41,500 candidates. Besides the estimated item parameters, that ranged from $a \in [0.26, 1.40]$, $b \in [-3.15, 2.51]$, and $c \in [0.00, 0.50]$, the calibration error was also reported in terms of standard deviations (sd a = 0.02, sd b = 0.044, sd c = 0.016). These standard deviations are relatively small, but that was expected because of the large sample of candidates. The larger the sample, the smaller the calibration errors.

Based on the estimated item parameters, the maximum amount of information over all theta levels (Hambleton and Swaminathan 1985, p. 107) was calculated for all items in the pool, and they were ordered from large to small. The information provided by the 50 most informative items is shown in Fig. 15.2.

Robust information could be calculated by subtracting the expected deviation for all of the items using Eq. (15.1).

A small experiment can show the impact of robust item pools. First of all, the deviation between the information provided by each item and its robust counterpart were calculated. Besides, for each item, three simulated counterparts were created by drawing item parameters from the multivariate normal distribution with a mean equal to the estimated item parameters and standard deviations equal to the calibration error. In this way, three simulated item pools were created. Deviations between the information provided by each item and its simulated counterparts were calculated as well. These deviations are shown in Fig. 15.3.

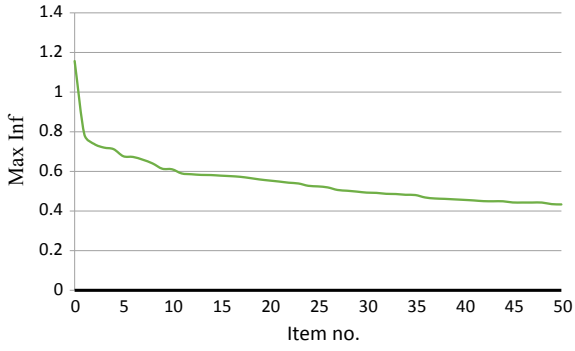

Fig. 15.2 Maximum amount of information provided by the 50 most informative items

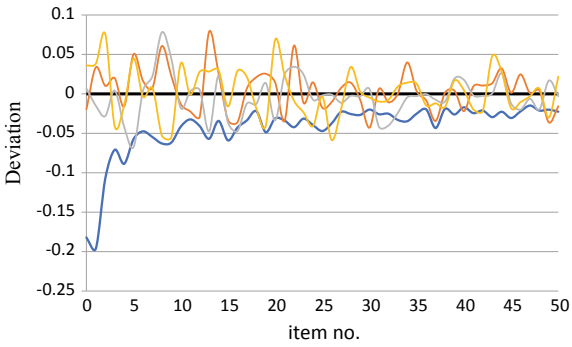

Fig. 15.3 Deviations from the maximum information for the robust information (thick line) and various simulated item banks (thin lines) for the 50 most informative items

From this experiment, several conclusions can be drawn. First of all, the robust counterparts provide less information that the original items. It should be noted however that the differences becomes smaller and smaller when the original items are less informative. Since the calibration errors differ for the various items, the deviance does not decrease monotonously. For the deviances between the original item and its simulated counterparts, it can be observed that the deviances are sometimes positive and sometimes negative. The most informative items had the largest calibration errors, therefore the largest deviations were observed for these items. Finally, it could be seen that simulated items could be less informative then their robust counterparts. The reason is that for most items in the middle of the ordered item pool, the robust counterparts are almost equal to the original items, even when the item parameters were estimated with considerable calibration error. This is in line with the observation that uncertainty was assumed to hit most for the most informative items.

15.3.3 Towards an Algorithm for Robust CAT

From this small experiment it can also be learned that for the 25 most informative items, the simulated items are more informative than their robust counterparts. In other words, the robust item information is still conservative. To deal with this conservatism, the Bertsimas and Sim method can be applied for item selection in robust CAT. This method assumes that uncertainty only affects the solution for at most Γ items in the test. The following pseudo-algorithm (Veldkamp 2012) describes the application of the Bertsimas and Sim method for selecting the gth item in CAT for a fixed length test of G items. It is a modified version or the original Bertsimas and Sim algorithm. In the first step, a robust item pool is used to calculate the conservative values d_i in the optimization model. Besides, the optimization model in Eqs. (15.3)–(15.6) has been formulated in such way that only the one item is selected that provides most information at the current ability estimate:

1. Calculate $d_i = I_i(\theta^{g-1}) - I_i^R(\theta^{g-1})$ for all items.

2. Rank the items such that $d_1 \geq d_2 \geq \ldots \geq d_n$
3. For $l = 1, \ldots, (G - (g - 1)) + 1$ find the item that solves:

$$G^l = \max\left\{\sum_{i=1}^{I} I_i(\hat{\theta}^{g-1})x_i - \left[\sum_{i=1}^{l}(d_i - d_l)x_i + \min(G - g, \Gamma)d_l\right]\right\} \quad (15.3)$$

subject to:

$$\sum_{i \in R^{g-1}} x_i = g - 1 \quad (15.4)$$

$$\sum_{i=1}^{I} x_i = g \quad (15.5)$$

$$x_i \in \{0, 1\} \quad i = 1, \ldots, I. \quad (15.6)$$

4. Let $l^* = \arg\max_{l=1,\ldots,n} G^l$.
5. Item g is the unadministered item in the solution of G^{l^*}.

In step 3 of the pseudo algorithm, (G-(g-1)) + 1 MIPs are solved, where (G-(g-1)) is the amount of items still to be selected. For the MIPs, it holds that x_i denotes whether item i is selected ($x_i = 1$) or not ($x_i = 0$) (see also Eq. [15.6]), and R^{g-1} is the set of items that have been administered in the previous $(g - 1)$ iterations. Equations (15.4)–(15.5) ensure that only one new item is selected. Finally, in (15.3) the amount of robust information in the test is maximized. This objective function consists of a part where the information is maximized and a part between square brackets that corrects for overestimation of the information. By solving (G-(g-1)) + 1 MIPs and choosing the maximum, a robust alternative for the test information that is not too conservative can be calculated. For details and proofs see Veldkamp (2013) and Bertsimas and Sim (2003).

15.4 Simulation Studies

To validate this algorithm for robust CAT, several simulation studies were conducted. The first study was conducted to illustrate the impact of item parameter uncertainty on CAT and to investigate whether robust item pools could reduce the effects. In the second study, the algorithm for robust CAT was studied. The Γ parameter, which indicates the number of items for which uncertainty is assumed to have impact on the resulting test, was varied to find out how this parameter influences the precision of the ability estimates. In the third study, the effects of five different methods for dealing with uncertainty in CAT were compared. First of all, we implemented Robust

CAT, where uncertainty in some of the items is assumed to impact ability estimation. The second method was more conservative. It is only based on the robust item pool, introduced in this chapter, where Fisher information of the items was corrected for expected uncertainty in the item parameters. In the second method, items were selected from the robust item pool. The third alternative was based on the work of Olea et al. (2012). They proposed to implement exposure control methods for dealing with uncertainty in the item parameters. Since exposure control methods limit the use of the most informative items, the use of the items with largest positive estimation errors will be limited as well. As a consequence, the impact of uncertainty in the item parameters on ability estimation will be neutralized. The fourth method combines Robust CAT and the exposure control method. Also because in practical testing situations, exposure control methods always have to be implemented to prevent that the most informative items in the pool become known (e.g. Sympson and Hetter 1985; van der Linden and Veldkamp 2004, 2007). Finally, the fifth alternative was to implement the Soyster (1973) method, where maximum values for the uncertainty for all the items was assumed. This method serves as a yardstick. It is very conservative, but takes all possible uncertainties into account.

15.4.1 Study 1

For the first simulation study, an item pool of 300 2PL-items was simulated, where the discrimination parameters a_i, $i = 1, \ldots, 300$, were randomly drawn according to $\log(a_i) \sim N\left(0, 0.3^2\right)$, and the difficulty parameters b_i, $i = 1, \ldots, 300$, were randomly drawn according to $b_i \sim N(0, 1)$. In this way, the true item parameters were simulated. Item parameter uncertainty was simulated by adding some random noise to these parameters according to $a_{ir} \sim N\left(a_i, (0.1)^2\right)$ and $b_{ir} \sim N\left(b_i, (0.3)^2\right)$. Test length was set equal to 20 items and items were selected based on maximum Fisher information. A number of 50,000 respondents were simulated for each $\theta \in \{-3, -2.75, \ldots, 3\}$. First, CATs were simulated based on the bank with uncertainty in the item parameters. Then, test information and RMSE were calculated based on the item parameters with uncertainty, true item parameters and based on robust item parameters.

15.4.2 Study 2

For the second study, the proposed method for robust CAT was implemented in R. Various settings of Γ were compared with the case where uncertainty was assumed to impact all items in the test. The item pool that was also used to illustrate the concept of robust item pools was applied. For this item pool, uncertainty in the parameter estimates was only small (average uncertainties in the parameters equal to $\Delta \mathrm{a} = 0.02$, $\Delta \mathrm{b} = 0.044$, $\Delta \mathrm{c} = 0.016$). To calculate the robust item pool, expected information

was calculated for all the items by taking only the uncertainty in the discrimination parameters into account (see also Veldkamp et al. 2013). In order to investigate the impact of the Γ parameter on CAT, uncertainty was assumed to impact the results for 25, 50, 75% and for all the items. This simulation study was much smaller than the first one. We simulated 1000 respondents for each of the ability values in the grid $(-3, -2.5, \ldots, 3)$. Test length was set equal to 20 items.

15.4.3 Study 3

In the third study, the five methods for dealing with uncertainty were compared with the Regular CAT, where uncertainty was not taken into account. In this study, also 1000 respondents were simulated for each of the ability values in the grid $(-3, -2.5, \ldots, 3)$. For the robust CAT method, Γ was set equal to 50% of the items. To study the impact of test length, the methods were compared for various test lengths. It varied from $n = 5$, $n = 10$, $n = 20$ to $n = 40$ items. In earlier studies on item selection in CAT (e.g. Matteucci and Veldkamp 2012) it turned out that differences between item selection methods only resulted in differences in ability estimates for short CATs with ten or fifteen items. The question remains whether the same findings hold for methods dealing with the impact of uncertainty on CAT.

15.4.4 Study Setup

Simulations for Study 1 were performed using dedicated software in C++, based on maximizing Fisher information and Warm's (1989) WLE estimator. Simulations for Study 2 and Study 3 were performed using the R software-package. The catR-package was used for implementing the CAT (Magis and Barrada 2017; Magis and Raîche 2012). In this package, several options are available. We applied the default settings with Bayes modal estimation, starting value equal to $\theta_0 = 0$ for all candidates, and a fixed test length as stopping criterion. The exposure control method in the package is based on Sympson-Hetter. To implement robust CAT in this package, we had to fix the number of items for which uncertainty had an impact in advance. In the robust CAT method, uncertainty in *at most Γ items* is assumed to impact the solution, but in the implementation, uncertainty in *exactly Γ items* was assumed to impact the solution. As a consequence, the robust CAT method became slightly more conservative.

15.5 Results

In Study 1, we compared CAT based on a robust item pool with CAT based on true item parameters and item parameters with uncertainty in them. Average test information functions are shown in Fig. 15.4.

The items in the robust item pool have been corrected for possible overestimation of the parameters. The resulting average information for CATs based on the robust item pool is lower than the information provided by CATs based on an item pool with uncertainty. In the middle of the ability distribution, the difference is only 2%, but towards the tails it is close to 10%. CATs were also simulated based on item parameters that were not disturbed by uncertainty. For these items it holds that they really provide most of their information when the theta estimated equals the difficulty parameter. Towards the tails of the distribution, there was quite a difference in average test information function. In the middle of the distribution, CATs are almost as or even more informative than CAT based on the disturbed item parameters. Root mean squared errors (RMSEs) for the various ability values are shown in Fig. 15.5.

Standard CAT, where uncertainty is not taken into account, resulted in an RMSE that is 6–17% higher than a CAT using the same item pool, but now with the item parameters assuming their true values. Thus, the efficiency of Standard CAT was

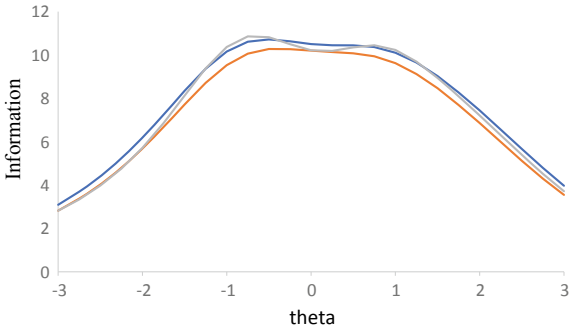

Fig. 15.4 Test information function for CAT with uncertainty in the item parameters (blue), robust item pool (orange) and based on real item parameters (green)

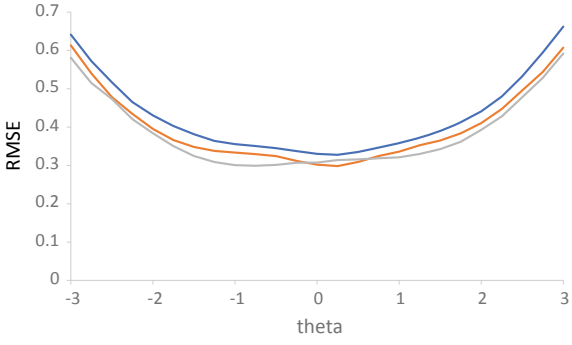

Fig. 15.5 RMSE for CAT with uncertainty in the item parameters (green), without uncertainty (blue) and based on robust item pool (blue)

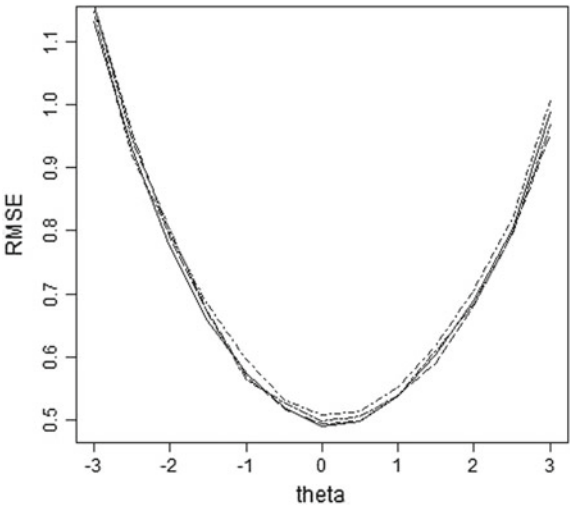

Fig. 15.6 RMSE for Robust CAT with $\Gamma = 25\%$ (solid line), $\Gamma = 50\%$ (large dashes), $\Gamma = 75\%$ (small dashes) and $\Gamma = 100\%$ (dash/dotted line) of the items

Table 15.1 Resulting average RMSEs for various methods for dealing with uncertainty for various test lengths (n)

Test length	n = 5	n = 10	n = 20	n = 40
Standard CAT	1.40	1.08	0.80	0.57
Robust CAT	1.40	1.09	0.82	0.59
Robust item pool	1.40	1.10	0.83	0.58
Exposure control	1.45	1.10	0.84	0.59
Robust item pool and exposure control	1.44	1.13	0.86	0.60
Soyster's method	1.49	1.15	0.88	0.63

overestimated by the same 6–17%. CAT based on robust item pools performed much better. RMSE was 5–9% higher than in simulations with perfectly known parameters, thus it overestimated the efficiency by 5–9%.

The second study focused on the method of robust CAT. In Fig. 15.6, the RMSE of the ability estimates is shown for various ability levels and various settings of Γ. The results for $\Gamma = 25\%$ (solid line) and $\Gamma = 50\%$ (large dashes) cannot be distinguished. For $\Gamma = 75\%$ (small dashes) the RMSE is slightly higher for abilities close to $\theta = 0$. For $\Gamma = 100\%$ (dash/dotted line), the RMSE is slightly higher for all ability levels. Overall, the differences in RMSE are very small.

The third study compared various methods for dealing with uncertainty in CAT. Impact of uncertainty was studied for various test lengths. Average RMSE was calculated over all ability values.

In Table 15.1, the results of various methods for dealing with uncertainty are shown for various test lengths.

Overall, it can be noticed that longer tests provide more information, and the differences between various methods in RMSE become smaller. Standard CAT resulted

in the smallest RMSE. Robust CAT performed only slightly worse. Robust item pools performed almost comparable to robust CAT. Both methods based on exposure control performed slightly worse. As expected, the combination of robust item pools and exposure control performed even worse than the exposure control method. Finally, Soyster's method, which is very conservative by nature, performed the worst. Some small deviances of this general pattern were noted, but this might be due to the relatively small sample size in this study.

15.6 Conclusion

In this chapter, the outline of a procedure for robust CAT was presented as an answer to the problem over capitalization on uncertainty in the item parameters in CAT. In this method, a robust item pool based on expected Fisher information and the robust item selection method of Bertsimas and Sim (2003) are combined. First, it was demonstrated how robust item pools can be used in CAT. In a large simulation study, it was illustrated that robust item pools can be implemented successfully, and that the resulting CATs are much closer to the real values than standard CAT that does not take uncertainty in the item parameters into account. Figure 15.6 illustrates how various implementations of robust CAT provide different results. $\Gamma = 100\%$ of the items is equivalent to selecting all the items from the robust item pool, where the other values of Γ only select a percentage of the items from this pool. The impact of Γ on the RMSE turned out to be small, but for $\Gamma \leq 50\%$ of the items, the best results were obtained. An explanation for the small impact of Robust CAT might be found in the construction of the robust item pool and the nature of CAT. In the robust item pool, expected information is calculated based on the assumption of a normally distributed estimation error. Large adaptations of the provided information are only made for small number of items. As was illustrated in Fig. 15.3, differences between Fisher information and robust item information are only small for most of the items. On top of that, only a few items will be selected per candidate where the robust item information is really much smaller than Fisher information due to adaptive item selection. Larger differences might be found in case of larger estimation errors in the item pool.

The method of Robust CAT was also compared with other methods for dealing with uncertainty in the item parameters in CAT. Robust CAT generally provided the smallest RMSEs. Only applying robust item pools, performed almost as well. Besides, the exposure control method did not perform that much worse. More conservative methods like the combination of a robust item pool with exposure control and Soyster's method had larger RMSEs. It should be remarked however, that the differences are relatively small.

All of these results were based on averages over large numbers of replications. It might be interesting to see what happens at the individual level. The Robust CAT method was developed to prevent overestimation of the precision at the individual level as well. The exposure control method, on the other hand, does not take over-

estimation at the individual level into account. For example, when the maximum exposure rate of the items is set equal to $r_{max} = 0.2$, this implies that items with overestimated discrimination parameters will still be used for 20% of the candidates. Especially for small CATs with test length smaller than 20 items, the impact might be considerable. Further research will be needed to reveal for which percentage of the candidates is affected.

Finally, it needs to be mentioned that Belov and Armstrong (2005) proposed using an MCMC method for test assembly that imposes upper and lower bounds on the amount of information in the test. Since there is no maximization step in their approach, item selection is not affected by the capitalization on chance problem. On the other hand, this approach does not take uncertainty in the item parameters into account at all. This could lead to infeasibility problems (Huitzing et al. 2005), as illustrated in Veldkamp (2013). Besides, MCMC test assembly was developed for the assembly of linear test forms, and therefore application to CAT is not straightforward.

References

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

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

Bertsimas, D., & Sim, M. (2003). Robust discrete optimization and network flows. *Mathematical Programming, 98,* 49–71.

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.

Eggen, T. T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing.* (Unpublished doctoral thesis, Enschede).

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.

Hambleton, R. H., & Swaminathan, H. (1985). *Item response theory, principles and applications*. Boston, MA: Kluwer Nijhoff Publishing.

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.

Lewis, C. (1985). *Estimating individual abilities with imperfectly known item response functions.* Paper presented at the Annual Meeting of the Psychometric Society, Nashville, TN.

Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, 76*(1), 1–19.

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software, 48*(8), 1–31.

Matteucci, M., & Veldkamp, B. P. (2012). On the use of MCMC computerized adaptive testing with empirical prior information to improve efficiency. *Statistical Methods and Applications.* (Online First).

Mislevy, R. J., Wingersky, M. S., & Sheehan, K.M. (1994). *Dealing with uncertainty about item parameters: Expected response functions* (Research Report 94-28-ONR). Princeton, NJ: Educational Testing Service.

Olea, J., Barrada, J. R., Abad, F. J., Ponsoda, V., & Cuevas, L. (2012). Computerized adaptive testing: The capitalization on chance problem. *The Spanish Journal of Psychology, 15,* 424–441.
Soyster, A. L. (1973). Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research, 21,* 1154–1157.
Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973–977).
Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*(2), 371–390.
van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer Verlag.
van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13,* 35–53.
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.
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.
Veldkamp, B. P. (2012). Ensuring the future of computerized adaptive testing. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in practice at RCEC* (pp. 35–46).
Veldkamp, B. P. (2013). Application of robust optimization to automated test assembly. *Annals of Operations Research, 206*(1), 595–610.
Veldkamp, B. P., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement, 37*(2), 123–139.
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450.