Chapter 4

Ensuring the Future of Computerized Adaptive Testing

Bernard P. Veldkamp

Abstract Capitalization on chance is a huge problem in computerized adaptive testing (CAT) when Fisher information is used to select the items. Maximizing Fisher information tends to favor items with positive estimation errors in the discrimination parameter and negative estimation errors in the guessing parameter. As a result, information in the resulting tests is overestimated and measurement precision is lower than expected. Since reduction of test length is one of the most important selling points of CAT, this is a serious threat to both the validity and viability of this test administration mode. In this chapter, robust test assembly is presented as an alternative method that accounts for uncertainty in the item parameters during test assembly.

Keywords: Automated test assembly, capitalization on chance, computerized adaptive testing, item parameter uncertainty, 0-1 LP, robust test assembly

Introduction

In computerized adaptive testing (CAT), item administration is tailored to the test taker. Tailoring the test turns out to entail a number of advantages. The candidate only has to answer items that are paired to his or her ability level, test length can be reduced, and test administration can be more flexible as a result of individualized testing. Besides, CATs could be offered continuously, on flexible locations, and even via the Web. The advantages of CAT turned out to be very appealing. Nowadays many CATs are run operationally in educational, psychological, and health measurement. Various algorithms for tailoring the test have been proposed. They 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. An update of the ability estimate is made after each administration of an item.
5. 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 assumptions underlying these CAT algorithms is that, for all items in the bank, the item parameters are known and can be treated as fixed values during test administration to calculate the amount of information provided. Unfortunately, this assumption is never met in practice. Item parameters have been estimated based on finite samples of candidates. The estimates might be unbiased, but they still have measurement error in them. This uncertainty is a source of concern. When test information is maximized, those items with high discrimination parameters will be selected from the bank. Positive estimation errors in the discrimination parameters will increase the amount of information provided, and therefore will increase the probability that the item will be selected. This phenomenon is also referred to as the problem of capitalization on chance.

Hambleton & Jones (1994) were among the first to study the effects of item parameter uncertainty on computerized construction of linear test forms from calibrated item banks. They found out that not taking the uncertainty into account resulted in serious overestimation of the amount of information in the test. Veldkamp (2012) illustrated this effect when he simulated an item bank of 100 items with uncertainty in them. All 100 items had the same parent, that is, all item parameters were drawn from the same multivariate distribution $N(\mu, \Sigma)$, with μ equal to the true item parameters $(a = 1.4, b = 0.0, c = 0.2)$ and Σ being the diagonal matrix with the standard errors of estimation $(SE\ a = 0.05, SE\ b = 0.10, SE\ c = 0.02)$. As a result the item parameters only varied due to uncertainty in the parameter estimates. Parameter ranges were $a \in [1.29, 1.52]$,

$b \in [-0.31, 0.29]$, and $c \in [0.14, 0.28]$. Ten items with highest Fisher information at $\theta = 0.0$ were selected from this bank for a test.

The resulting test information function was compared to the test information function based on the true item parameters $(a = 1.4, b = 0.0, c = 0.2)$. As can be seen in Figure 1, the test information is overestimated by 20%, when uncertainty is not taken into account.

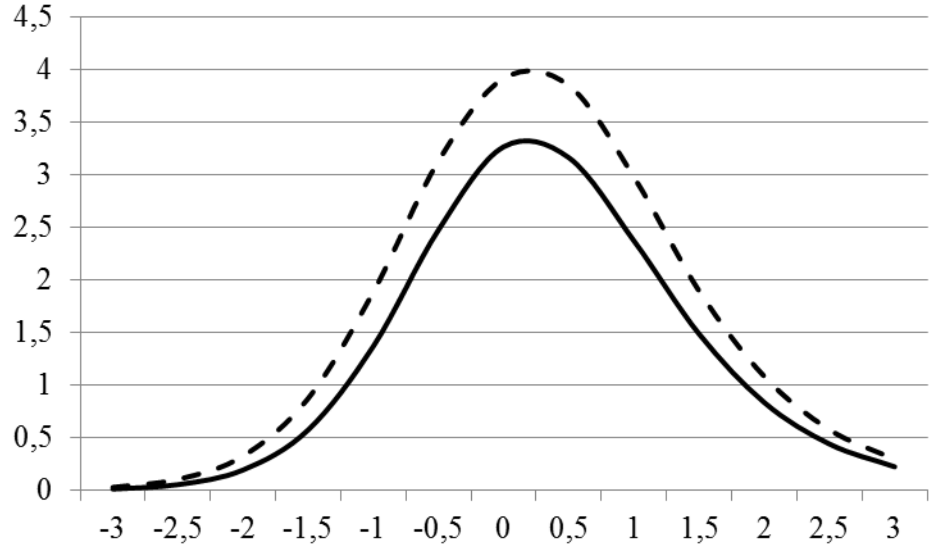

Figure 1 Test information function: ATA (dashed line) or true (solid line)

Hambleton & 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 applied to CAT, calibration sample size plays a comparable role. The ratio of item bank size to test length is more of an issue in CAT, since only one item is selected at a time, which results in an even less favorable ratio. Olea, Barrada, Abad, Ponsoda, & Cuevas (2012), 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 huge problem in CAT when Fisher information is used to select the items. The measurement precision of the test is vastly overestimated. Alternative strategies for item selection in CAT will have to be used so as not to compromise the validity of this test administration mode.

Robust Test Assembly

In combinatorial optimization, mathematical techniques are applied to find optimal solutions within a finite set of possible solutions.

The set of possible solutions is generally defined by a set of restrictions. Automated test assembly (ATA) problems are a special case of combinatorial optimization problems. The objective of ATA is often to maximize the information in the test, and the set of possible solutions is generally defined by the test specifications, for example, by the content constraints. An extensive introduction to the topic of formulating ATA problems as mixed integer programming (MIP) problems can be found in van der Linden (2005).

To solve the problem of dealing with uncertainty in the item parameters in CAT, a first step would be to search the literature for methods that have been proposed to deal with parameter uncertainty in combinatorial optimization. Soyster (1973) was among the first to present a method for dealing with uncertainty in combinatorial optimization problems. He assumed that for every uncertain parameter an interval could be defined that contained all possible values. He replaced each uncertain parameter by its infimum and solved the problem. This solution served as a robust lower bound for the solution of the original problem. Unfortunately, this method was very conservative. It assumed a maximum error in all the parameters, which is highly unlikely in practice. The good thing, however, was that Soyster (1973) opened up a new area of research: robust optimization. The ultimate goal of robust optimization (Ben Tal, El Ghaoui, & Nemirovski, 2009) is to take data uncertainty into account when the optimization problem is solved in order to "immunize" resulting tests against this uncertainty. Under this approach, a suboptimal solution is accepted in order to ensure that the solution remains near optimal when the estimated parameters turn out to differ from their real values. For ATA this means that uncertainty in the item parameters or in the information function is taken into account during test assembly to immunize the test against overestimation of the test information.

De Jong, Steenkamp, & Veldkamp (2009) applied a modified version of Soyster's method to ATA,when they constructed country-specific versions of a small marketing scale. Instead of replacing uncertain parameters by their infima, they subtracted one posterior standard deviation from the estimated Fisher information as a robust alternative. Veldkamp, Matteucci, & de Jong (2012) studied this modified Soyster method in more detail, for example, they studied differences in effects of uncertainties in various item parameters in test assembly.

Veldkamp (2012) studied a different approach based on the robust optimization method developed by Bertsimas and Sim (2003). Instead of doing a small correction (minus one standard deviation of the uncertainty distribution) for all items in the bank, a substantive correction (replacing the parameters by their infima) is made only for the maximum number of items assumed to affect the solution.

This resembles more closely the practice of ATA, where some items in the test will have high positive estimation errors, while others will not. A robustness level Γ (i.e., the maximum number of item parameters that might be replaced) has to be defined beforehand. Γ can vary anywhere from zero (which resembles ATA) to all items in the test (which resembles the Soyster method). When the ratio of item bank size to test length is small, many items will be selected from the item bank. Γ will be close to zero, because only a few of the selected items will have high positive estimation errors. When the ratio of item bank size to test length is high, only a very small proportion of the items will be selected from the bank. Capitalization of chance will be more of an issue, and Γ will be closer to the test length. Bertsimas and Sim (2003) proved that finding an optimal solution for a combinatorial optimization problem where at most Γ parameters were allowed to change, was equal to solving $(\Gamma+1)$ MIP problems. For details of the method, see Veldkamp (2012).

Robust CAT Based on Expected Information

Even though relatively 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, 2012), both methods do not use information known about the distribution of the item parameter uncertainty. Uncertainty in the item parameters results from parameter estimation, and it is assumed to follow a normal distribution with a mean equal to the parameter estimates and a standard deviation equal to the standard error of estimation for maximum likelihood estimation, or to the posterior standard deviation in a Bayesian framework. This information could be used to calculate the expected information for each item, taking the uncertainty distribution of the parameters into account. Lewis (1985) already proposed using expected response functions (ERFs) to correct for uncertainty in the item parameters (Mislevy, Wingersky, & Sheehan, 1994) for fixed-length linear tests. The same idea might be applied at the item bank level as well, thus providing a starting point for a robust test assembly procedure for CAT.

Robust Item Pool

The first step in such a procedure would be to develop a robust item pool. Since the uncertainties in the parameters are assumed to follow a normal distribution, the cumulative distribution function can be used to calculate which percentage of the items is expected to have which deviation. For example, 2.5% of the items are assumed to have a positive deviation larger than 1.96 standard deviations.

Based on this information, robust item information can be calculated by subtracting the expected deviation from the estimated item information. When all items in the bank are ordered from smallest to largest with respect to their maximum information, the robust item information can be calculated as:

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

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. Within a Bayesian framework, a comparable procedure has to be applied, where the posterior distribution is used to calculate z_i.

Empirical Example

To illustrate the effects of expected information, robust item information was calculated for all items of an operational item bank. 306 items were calibrated with a three-parameter logistic model (3PLM):

$$P_i(\theta) = c + (1-c)\frac{e^{a(\theta-b)}}{1+e^{a(\theta-b)}}, \tag{2}$$

where a is the discrimination, b is the difficulty, and c is the guessing parameter. The item parameters were estimated using BILOG MG 3, for a sample of 41,500 candidates. The estimated parameter ranges were $a \in [0.26, 1.40]$, $b \in [-3.15, 2.51]$, and $c \in [0.00, 0.50]$, and the average uncertainties were $(\Delta a = 0.02, \Delta b = 0.044, \Delta c = 0.016)$.

The maximum amount of information over all theta levels (Hambleton, & Swaminathan, 1985, p.107) provided by the 50 most informative items is shown in Figure 2.

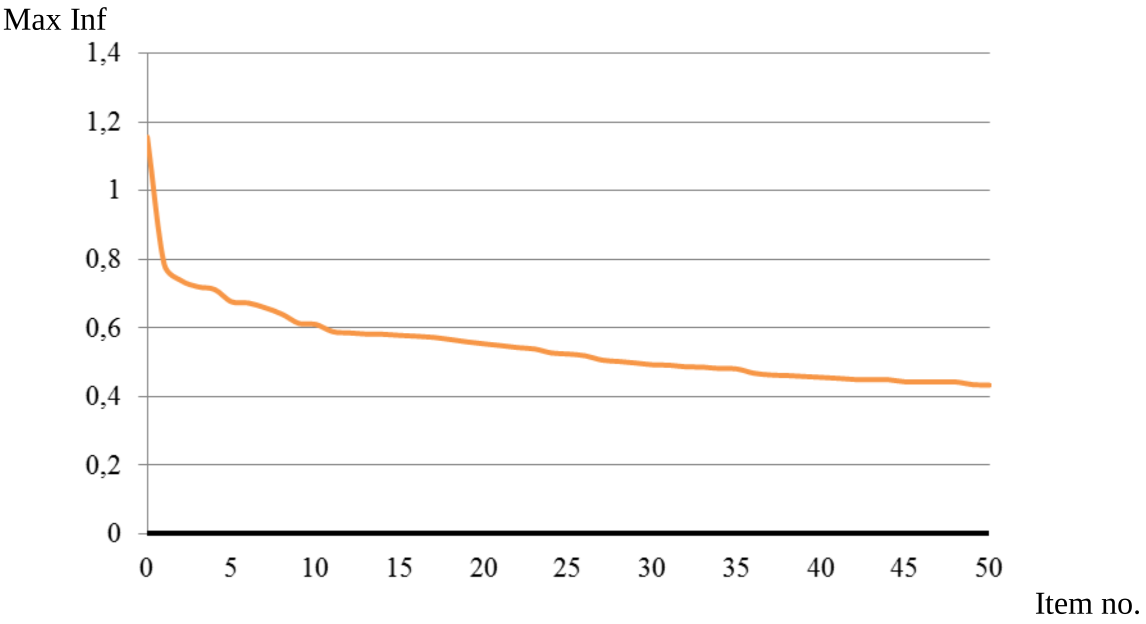

Figure 2 Maximum amount of information provided by the 50 most informative items

All items were ranked with respect to their maximum amount of information over all theta levels, and the robust information was calculated by subtracting the expected deviation for all of the items. To illustrate how robust item information corrects for uncertainty in the item parameters, its performance was compared with a number of simulated item banks. Three item banks were simulated by randomly drawing item parameters from the multivariate normal distribution with a mean equal to the estimated item parameters and standard deviations equal to the errors of estimation. The deviance in maximum information between the estimated item parameters on the one hand and the robust and simulated item parameters on the other hand is shown in Figure 3.

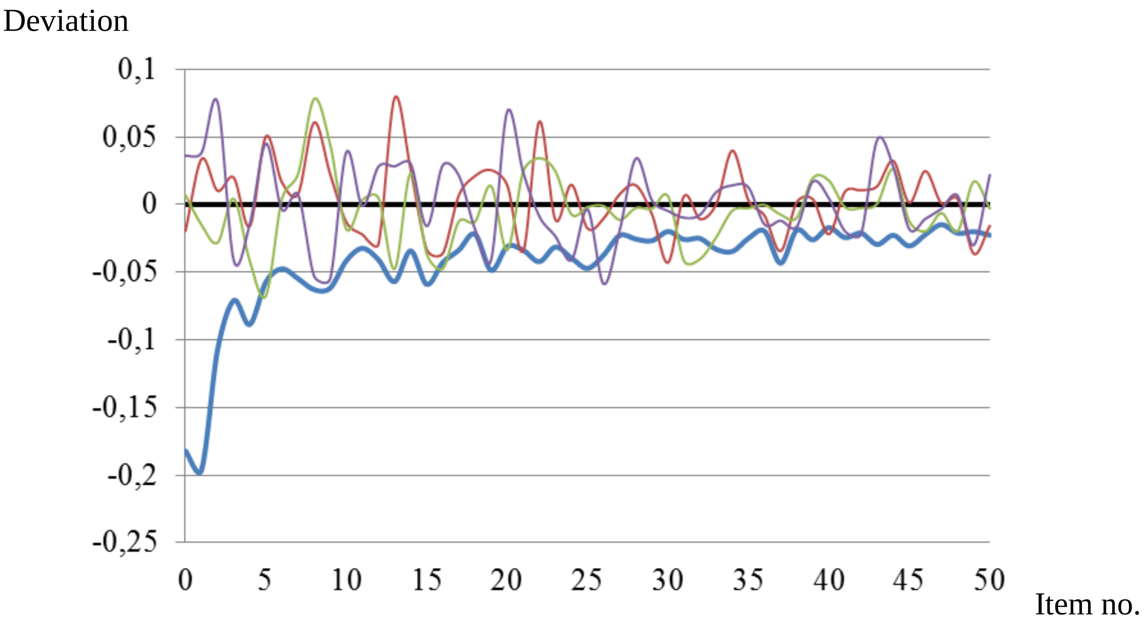

Figure 3 Deviations from the maximum information for the robust information (thick line) and various simulated item banks (thin lines) for the 50 most informative items

As expected, the robust maximum information is generally smaller than the estimated maximum information for the 50 most informative items, but the difference becomes smaller and smaller when the items are less informative. Because of the differences in $SD(I(\theta))$ for the various items, the robust maximum information does not increase monotonically. As can be seen in Figure 3, $SD(I(\theta))$ for the second item is larger than $SD(I(\theta))$ for the first item. The curves of the deviances for the simulated item banks hover around zero. By chance, the deviation will be positive for some of the items and negative for others. It can also be seen that for individual items, the deviance for the simulated information could even be larger than the deviation of the robust information, but for a test, which is for a group of items, the robust maximum information serves pretty well as a lower bound.

Robust Item Selection

The robust item information is still conservative. It assumes that uncertainty hits where it hurts most; that is, it assumes that the most informative items have the highest uncertainty in them. In practice, however, this is not the case. This can also be seen in Figure 3, where for the first 25 items, the robust maximum information is obviously smaller than the simulated maximum information. To correct for this conservatism, the Bertsimas and Sim method can be applied for item selection in the second step of robust CAT.

This method assumes that uncertainty only affects the solution for at most Γ items in the test. The following pseudo-algorithm describes the application of the Bertsimas and Sim method for selecting the g th item in CAT for a fixed length test of G items:

1. Calculate $d_i = I_i(\theta^{g-1}) - I_i^R(\theta^{g-1})$ for all items.
2. Rank the items such that $d_1 \geq d_2 \geq ... \geq d_n$
3. For $l = 1,...,(G-(g-1))+1$ find the item that solves:

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

subject to:

$$\sum_{i \in R^{g-1}} x_i = g-1 \tag{2}$$

$$\sum_{i=1}^{I} x_i = g \tag{3}$$

$$x_i \in \{0,1\} \quad i = 1,...,I. \tag{4}$$

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

In step 3 of the pseudo algorithm, (G-(g-1))+1 MIP 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 Equation [6]), and R^{g-1} is the set of items that have been administered in the previous $(g-1)$ iterations.

Equations (4)–(5) ensure that only one new item is selected. Finally, in (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.

This correction term varies for each value of $l = 1,...,(G-(g-1))+1$. d_l represents the overestimation of the information in item *l*. When $l = 1$, d_l is equal to the largest overestimation of item information at the estimated ability level in the item bank, and Γ times d_1 (or (G-(g-1)) d_1, when less than Γ items are remaining) is subtracted as a correction. This will be too conservative because there is only one item with the maximum overestimation of the information. For larger values of l, the amount of overestimation is smaller, which implies that the correction factor is smaller , and the solution is less conservative.

For these values of *l* it is taken into account that selecting one of the items with *i*<*l* results in a larger overestimation, since, as a result of the ordering in step 2, $d_i > d_l$. 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 (2012) and Bersimas & Sim (2003).

Conclusion and Discussion

Capitalization on chance is a serious problem in CAT that might negatively affect both the validity and viability of this test administration mode. In this chapter, the outline of a procedure for robust CAT was presented as an answer to this problem. It accepts a suboptimal solution that remains near optimal even when item parameters turn out to be seriously overestimated. The next step in this research would be to carry out an extensive simulation study to determine its strengths and weaknesses.

Other methods have been proposed in the literature to deal with the problem of capitalization on chance. 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, Veldkamp, & Verschoor, 2005), as illustrated in Veldkamp (2012). Besides, MCMC test assembly was developed for the assembly of linear test forms, and therefore application to CAT is not straightforward.

Olea et al. (2012) propose using item exposure control to deal with this problem. When items are selected based on maximum information, the most informative items tend to be selected more often than the others. Exposure-control methods can be implemented to limit item exposure and force less informative items to be selected. In this way, selection of the most informative items due to capitalization on chance will be prevented. Olea et al. (2012) report some promising results. Instead of correcting for uncertainty, this method limits the probability that items most vulnerable to overestimation of their information will be selected. A combination of robust CAT and item exposure control would probably result in a very strong method to prevent the capitalization on chance problem in CAT.

Every operational CAT program seriously needs to consider the impact of uncertainty in the item parameters on the reported measurement precision. Various simulation studies by Hambleton and Jones (1994), Olea et al. (2012), Veldkamp (2012), and Veldkamp et al. (2012) reported overestimation of the amount of information in the test of up to 40%. When the uncertainty in the item parameters is known, simulation studies have to be carried out to determine the impact on the specific CAT program at hand. Once the impact is known, one can decide either to neglect the problem or to implement a method that deals with item parameter uncertainty either implicitly (by applying exposure-control methods) or explicitly by using robust CAT, or by a combination of both.

Acknowledgement

This chapter was written when the author was a visiting scholar at the Law School Admission Council (LSAC). The opinions and conclusions contained in the chapter are those of the author and do not necessarily reflect the positions or policy of LSAC.

References

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.

Bertsimas, D., & Sim, M. (2003). Robust discrete optimization and network flows. *Mathematical Programming, 98,* 49–71, DOI:10.1007/s10107-003-0396-4

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

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

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. DOI:10.1111/j.1745-3984.2005.00012.x

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

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. DOI:10.5209/rev_SJOP.2012.v15.n1.37348

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

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

Veldkamp, B. P. (2012). *Application of robust optimization to automated test assembly* (Research Report 12-02). Newtown, PA: Law School Admission Council.

Veldkamp, B. P., Matteucci, M., & de Jong, M. (2012). *Uncertainties in the item parameter estimates and robust automated test assembly.* Manuscript submitted for publication.