LSAC RESEARCH REPORT SERIES

■ Detection of Advance Item Knowledge Using Response Times in Computer Adaptive Testing

Rob R. Meijer
Leonardo S. Sotaridona
University of Twente, Enschede, The Netherlands

■ Law School Admission Council
Computerized Testing Report 03-03
May 2006

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are over 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

The shift in test delivery from traditional paper-and-pencil (P&P) testing to computerized adaptive testing (CAT) was motivated by two main goals: increasing measurement efficiency and improving item security. Additional practical advantages of CAT are: the possibility of easier and more frequent test administration; immediate feedback of results; increased test taker interest; enhanced measurement precision; and reduced cost of test production, administration, and scoring.

Though item security was initially believed to be one of the greatest advantages of CAT, it soon became one of its major problems. To ensure item security, item banks needed to be continually updated. This necessity greatly increased the cost of implementing an operational CAT. A number of approaches to tackle the challenge of ensuring item security have been proposed.

In this paper, the use of information obtained from response times for detecting item security problems is investigated. We propose a new method based on an estimate of the "effective response time" for test takers on each item, which is based upon a method previously proposed by other researchers. Effective response time is defined as the time required for a test taker to answer an item correctly. An unusually short response time relative to the expected effective response time may be an indicator of item preknowledge. The test statistic in this new method was applied in an empirical study. Results showed that the false-positive error rate of the statistic could be controlled. Further analysis revealed that the detection rates produced by the statistic are sensitive to the magnitude of the reduction in response time as a result of item preknowledge.

Abstract

We propose a new method for detecting item preknowledge in a CAT based on an estimate of "effective response time" for each item. Effective response time is defined as the time required for an individual examinee to answer an item correctly. An unusually short response time relative to the expected effective response time may be an indicator of item preknowledge. The new method was applied to empirical data. Results showed that the Type I error rate of the statistic can be controlled. Power analysis revealed that the power is high when the response time is reduced even for a small set of items where the examinee has item preknowledge.

Introduction

Recent advances in psychometrics and the development of fast computing facilities enabled the development of computerized testing. The shift in test delivery from traditional paper-and-pencil (P&P) testing to computerized testing is motivated, among others, by two main goals: test efficiency and test security. Computer-based testing may be categorized into several different forms (Parshall, Spray, Kalohn, & Davey, 2002), namely, linear testing, linear-on-the-fly testing, stratified testing, classification testing, and adaptive testing, where linear testing is the least efficient and least secure form of computer-based testing and adaptive testing is the most efficient and the most secure form. Computerized adaptive testing (CAT) is efficient because the most appropriate items for a test taker are selected, which results in shorter tests with the same measurement precision as longer P&P tests.

Additional advantages of CAT are (Meijer & Nering, 1999): ease and frequency of administration; immediate feedback of results; increased examinee interest; and reduced cost of test production, administration, and scoring. A disadvantage of CAT is item exposure (Steinberg, 2002). Although test security initially seemed to be one of the greatest advantages of CAT, it became one of its major problems. Item banks needed to be continually updated to ensure item and test security. This greatly increased the cost of implementing an operational CAT. A number of approaches have been proposed to solve this problem. Methods used to deal with item exposure (Meijer & Nering, 1999) are (1) controlling item exposure rates during administration (Sympson & Hetter, 1985), (2) managing item banks, and (3) examining and correcting any unusual responses after the test has been administered. The goal of exposure control procedures is to limit item usage by limiting the frequency with which items are administered. Different procedures have been proposed such as the random-from-best-n method (e.g., Kingsbury and Zara, 1989, p. 369–370), the count-down random method (e.g., Stocking and Swanson, 1993, p. 285–286), and the multinomial procedure (Stocking & Lewis, 1995), which controls the frequency of reference to a particular target population of test taker ability. Item bank management is comprised of procedures such as bank rotation (e.g., van der Linden, Veldkamp, & Reese, 2000).

In this study, we will deal with methods that examine unusual responses after the test has been administered. Several CAT studies have dealt with this topic. For example, McLeod and Lewis (1999) investigated detection of item preknowledge and item memorization in a CAT environment, and Veerkamp (1996) and van Krimpen-Stoop and Meijer (2000) investigated previously known items in CAT and person

misfit,
respectively, using statistical process control techniques. A limitation of some of these methods is the low detection rates and false-alarms rates that are too high to be useful in practice.

One of the advantages of a CAT is that non-conventional item formats are used and that response times are often registered. Response times provide information for investigating item preknowledge. Several models for response times are available; see, for example, Thissen (1983) and Schnipke & Scrams (1997). Van der Linden and van Krimpen-Stoop (2003) proposed a method where response times are used to detect aberrant responses in computerized adaptive testing. In this study we will adapt their method to detect item preknowledge.

Before we present a method for detecting preknowledge of items or cheating on a CAT, it should be noted that methods that were developed to detect cheating in P&P tests cannot be applied to CAT. The nature of cheating in P&P tests is different from cheating in CAT. In P&P tests, the suspected cheater may be someone who copies answers from another examinee while in CAT, answer copying is less common since each examinee is seated such that copying is almost impossible. Furthermore, in CAT, each examinee receives different items. While answer copying is a common type of cheating in P&P tests, item preknowledge is a dominant form of cheating in CAT. Items administered to each examinee are selected such that the item difficulty level matches the current estimate of the examinee ability. As a result, the length of the test is reduced and the probability of responding correctly to the item is around .5. Methods developed for P&P tests that are based on the difference between observed and expected response (e.g., ω, Wollack, 1997) will have low power to detect cheaters in CAT. Furthermore, since the examinees received different sets of items, methods that are based on matching answers between the two examinees (e.g., K-index, Holland, 1996; g_2, Frary et al., 1977; S_1 and S_2, Sotaridona & Meijer, 2002; 2003) cannot be used in CAT.

It should be noted that the approach described here may be a helpful tool for identifying unexpected response behavior as a result of advanced knowledge of items, but it is by no means a *sufficient* tool for accusing an examinee of advanced knowledge of items. We see this research as one piece of evidence that may alert researchers that there is “something going on.” It can be used together with other kinds of evidence to build up a case related to unexpected response behavior.

This study is organized as follows. First, we will discuss a method for estimating the effective response time for each item, where large positive deviations between the observed and the expected response times are used as indicators of item preknowledge. Second, we will investigate the statistical properties of this method using empirical data from a CAT. Third, we will investigate the power of these methods to detect item memorization.

Effective Response Time

The “effective response time” (ERT) is the time an individual examinee *j* with an ability level θ_j uses to answer an item *i* correctly. The idea is to establish the ERT for each item for each examinee *j*. To do this, (1) examinee *j* should be able to answer item *i* correctly, that is, the probability of a randomly selected examinee *j* answering an item correctly, P_{ij}, should be large enough, and (2) examinee *j*’s response to item *i* should be correct. The reasons for these two requirements are explained below. The effect of item preknowledge is that examinee *j* answers item *i* correctly. If we can establish an accurate estimate of ERT, examinee *j*, who answered item *i* correctly as a result of item preknowledge, is expected to have a response time deviating from the ERT. Thus, we want to distinguish cheaters from non-cheaters through their response time.

The rationale behind the two requirements outlined above is to reduce the variability in the observed response time due to other types of answering behavior. For example, a less able examinee may answer an item correctly by guessing without spending much time considering the item, leading to an observed response time that is unlikely short. Some examinees may guess at the answer to an item without even reading it, and hence the observed response time will be very low. The item is too difficult for the examinee and the probability of getting the item correct, P_{ij}, is small. The observed response time for these examinees may not reflect the true or effective response time that is expected in order to answer an item correctly. On the other hand, there may be examinees who spend a great deal of time on an item they are not capable of answering and still record an incorrect response. In all these situations, the observed response time is uninformative, or even misleading, for estimating the ERT. Thus, less able examinees are excluded in estimating the ERT time because, by definition, ERT is the time required by an *able* examinee to correctly answer the item *i*. Less able examinees may spend less time working on item *i*, and guess the correct answer to an item. The response time of these examinees does not contribute to the estimation of ERT. In fact, the effect is the opposite, that is, it adds more variability or noise to the data. In principle, cheaters may artificially spend some reasonable time on item *i* to eliminate suspicion. This strategy might be very difficult to remedy or discover. On the other hand, detection through ERT must be combined with other evidence such as deviation between the observed and expected response.

Another possibility is that an examinee could respond incorrectly due to carelessness or because he/she is not motivated. Although these examinees may also answer an item correctly in a shorter amount of time by randomly guessing or using partial information to determine the correct answer, we still want to include them because potential cheaters or examinees who are using an item preknowledge strategy could be identified from these groups. Another reason to focus on examinees who answered an item correctly is that an examinee who knows an item a priori will answer the item correctly.

To eliminate the effect of item characteristics in estimating ERT, the ERT is estimated independently for each item. By modeling each item independently, the result is independent of any particular characteristics of an item. For example, an easy item may contain a longer reading passage so that the time required to answer the item may be longer than that for a more difficult item with a shorter reading passage.

A Model for Item Response and Response Time

The answering behavior of a person is often described by the 1-, 2-, or 3-parameter logistic models (l, 2, 3 PLM; van der Linden & Hambleton, 1997). The 3PLM is given by

$$P_{ij} \equiv \Pr(U_{ij} = 1) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]} \tag{1}$$

where U_{ij} is a response indicator of examinee j to item i (1 if correct and 0 if incorrect), and $a_i \in (0,\infty)$, $b_i \in R$, and $c_i \in [0, 1)$ are the discrimination, difficulty, and guessing parameters for item i, respectively. Setting $c_i = 0$, we have the 2PLM, and if $c_i = 0$ and $a_i = 1$, we have the 1PLM or the Rasch model.

As discussed in van der Linden & van Krimpen-Stoop (2003), a loglinear model can be used to model the response time

$$\ln T_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij} , \tag{2}$$

with

$$\varepsilon_{ij} \sim N(0,\sigma^2) , \tag{3}$$

where δ_i is a parameter for the response time required for item i, τ_j is a parameter for the slowness of examinee j, μ is a parameter indicating the general response time level for the population of examinees and pool of items, and ε_{ij} is a normally distributed residual or interaction term for item i and examinee j with mean 0 and variance σ^2. It follows that $\ln T_{ij} \sim N(\mu + \delta_i + \tau_j, \sigma^2)$. The parameters of Equation 2 can be estimated as follows:

$$\mu \equiv E_{ij}(\ln T_{ij}) \tag{4}$$

$$\delta_i \equiv E_j(\ln T_{\mathrm{ij}}) - \mu \tag{5}$$

$$\tau_j \equiv E_i(\ln T_{\mathrm{ij}}) - \mu . \tag{6}$$

Proposed Approaches

We assume that the item parameters are known, that is, they are estimated during the calibration stage. Given fixed item parameters, the ability of the examinees is estimated based on the data. Given the item parameters and $\hat{\theta}_j$, P_{ij} is computed using Equation 1. From the examinees who were presented a specific item i, an examinee $j = 1, ..., J_i$ is selected such that $P_{ij} > \gamma$, and $U_{ij} = 1$. Note that for each item, there will be different sets of examinee j because in CAT, not all examinees respond to the same items.

First, an estimate of τ_j is obtained using Equations 4–6. The effective response time is modeled as

$$\ln T_{ij} = \beta_0 + \beta_1\theta_j + \beta_2\tau_j + \varepsilon_j , \tag{7}$$

where θ_j and τ_j are treated as known regressors, the β's are regression coefficients, and ε_j is an error term assumed to be normally distributed with mean 0 and variance σ_i^2. The formulation in Equation 7 suggests that each item is modeled independently of the other items. By modeling each item independently, the effect of item characteristics on response time is eliminated and this yields a more accurate estimate of effective response time. Observed response times that are significantly lower than expected can be used as evidence of item preknowledge. We assume that the response time is normally distributed in the log scale. It follows that

$$\widehat{\ln T_{ij}} \equiv E(\beta_0 + \beta_1\theta_j + \beta_2\tau_j + \varepsilon_j) = \hat{\beta}_0 + \hat{\beta}_1\theta_j + \hat{\beta}_2\tau_j \ . \tag{8}$$

Let c be an examinee index for the suspected copier or examinee suspected of having item preknowledge. The response time of examinee c to an item is evaluated against the expected response time for that item. We assume that the response time is normally distributed in a log scale. Then it follows that

$$z_{ic} = \frac{\ln T_{ic} - \widehat{\ln T_{ij}}}{\sigma_i} \tag{9}$$

is standard normal, where

$$\sigma_i^2 = (J_i - 1)^{-1} \sum_j^{J_i} (\ln T_{ij} - \widehat{\ln T_{ij}})^2 \tag{10}$$

is the variance of the log response time for item i. The variate z_{ic}^2 is chi-squared distributed with one degree of freedom. Hence, the statistics

$$X_c = \sum_i z_{ic}^2 \sim \chi_{Ic}^2 \tag{11}$$

can be used to detect item preknowledge. For example, the quantity:

$$\Pr\,(X_c \geq x) = p \tag{12}$$

can be compared to some level of significance α. The value of p that is less than α is indicative of item preknowledge.

If there is not much variability in the observed response times of examinees on item i, one may replace $\widehat{\ln T_{ij}}$ in Equation 8 by the estimated mean

$$\overline{\ln T_{ij}} = J_i^{-1} \sum_j^{J_i} (\ln T_{ij}) \ . \tag{13}$$

Figure 1 shows the scatter plots of log response time across θ for four items. The scatter plots in the first column in Figure 1 include all examinees who were administered the item whereas those in the second column include only those examinees who answered the item correctly and for whom $P_{ij} \geq \gamma$. We note that the log response times are more stable in the second column than in the first column, which indicates that the conditioning had the effect of stabilizing the variability of the observed response times.

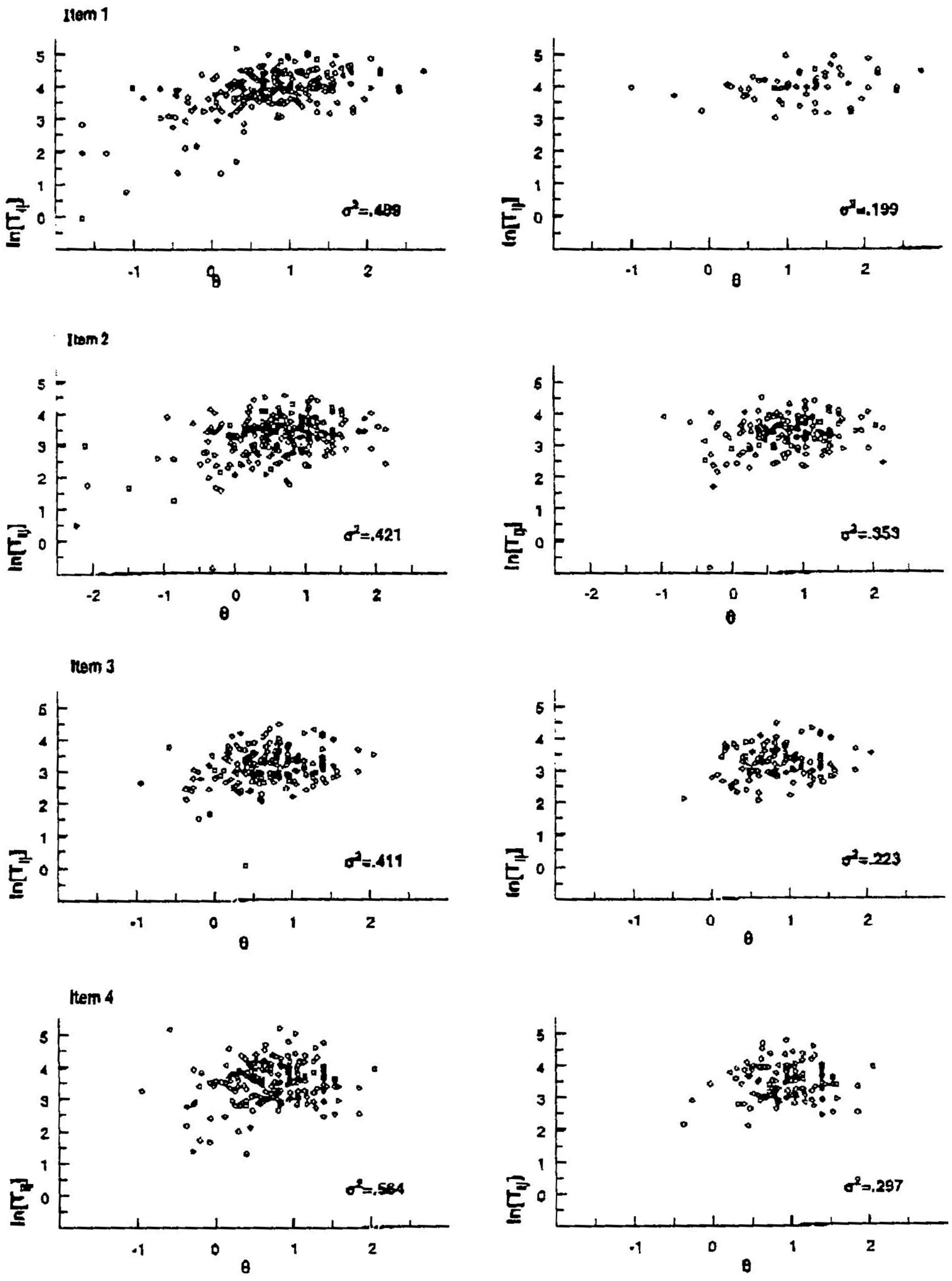

FIGURE 1. *Scatters of log response time (ln[Tij]) across θ. First column includes all examinees who were administered the items. Second column includes only examinees who answered correctly and had* $P_{ij} \geq \gamma$.

We refer to an approach based on Equation 9 as Approach 1. If the second term of Equation 9 is replaced by $\overline{\ln T_i}$, we refer to it as Approach 2.

The test in Equation 12 is a one-sided test since the likely effect of item preknowledge is to reduce the response time. Note that two different observed response times $t_1 = t - \bar{t}$ and $t_2 = t + \bar{t}$ for the same examinee will yield the same value of z^2 but only t_1 is informative for detecting item preknowledge. To minimize the Type I errors due to the effect of t_2, it is necessary that the items in Equation 11 are such that $\ln t_{ij} \leq \widehat{\ln t_{ij}}$ or $\ln t_{ij} \leq \overline{\ln t_{ij}}$.

Method

Using real data, we will investigate the Type I error of the proposed approaches. We will also conduct a small simulation study to investigate the detection rate of the two approaches.

The Data

Item score patterns of 528 examinees were available from an abstract reasoning test (ART) administered to the freshmen of the University of Kansas. The test was designed as a computerized adaptive test. The ability parameter was estimated using expected a posteriori (EAP) and the one-parameter logistic model.

Type I Error and Detection Rates

One hundred examinees were selected randomly (without replacement) from the 528 examinees who took the computerized adaptive test on abstract reasoning. The probability in Equation 12 was computed for each examinee and then compared to $\alpha = .01$ and .05. The process was replicated 100 times and the mean proportion of examinees with p less than or equal to α was used as an estimate of the Type I error. Note that we assumed there were no large percentages of persons in this sample with advance knowledge of the items. This was a reasonable assumption because the test was especially developed for the University of Kansas and was not commercially available; strict confidentiality rules were also taken into account.

To investigate the power of the methods, a random sample of examinees from the real data set was selected and their response time was changed to one-half or one-fourth of the original response time on one-half or three-fourths of all the items they responded to. These examinees constitute the simulated examinees having item preknowledge. The detection rate was computed as the proportion of 1,000 simulated cheater examinees having p less than or equal to α. The 1,000 examinees were the result of resampling 1,000 times from the original sample.

Results

Type I Error and Detection Rates

For nominal $\alpha = .05$, we found an empirical $\alpha = .022$ (Approach 1) and $\alpha = .038$ (Approach 2); for $\alpha = .01$, we found an empirical $\alpha = .009$ (Approach 1) and $\alpha = .011$ (Approach 2). It can be concluded that Approach 1 is more conservative than Approach 2, especially for $\alpha = .05$. For $\alpha = .01$, the Type I errors of both approaches are very near the nominal levels.

In Table 1, it can be seen that the detection rate of Approach 2 is higher than the detection rate of Approach 1. This is expected since Approach 1 is more conservative than Approach 2. Both approaches are sensitive to the amount of time reduced due to item preknowledge. For example, both approaches have high power to detect item preknowledge for cheaters who know at least half of the items and whose quick response is equal to one-fourth of the normal time.

TABLE 1
Detection rates

α	.05	.01
$n = .50$, $r = .50$	.345 (.475)	.216 (.346)
$n = .50$, $r = .25$	.938 (.944)	.878 (.895)
$n = .75$, $r = .50$	.482 (.585)	.311 (.447)
$n = .75$, $r = .25$	.976 (.985)	.949 (.965)

n = proportion of items known

r = proportion of the time reduction; for example, $r = .25$ means that if the normal time is 80 seconds, the new time is set to .25 of 80 and that is 20 seconds

Discussion

The proposed approach based on the ERT seems sensitive to identify item preknowledge in CAT. When we used all the data (i.e., the response times of all examinees who were administered item i), the variability of the observed response time was larger than when only those examinees who answered the item correctly and whose $P_{ij} \geq \gamma$ (e.g., $\gamma = .25$) were included. For example, we noted that the ln T_{ij} is almost constant across θ. This suggests that the simple average will do instead of using regression to estimate the ERT of an item i.

Results of the empirical and simulation study showed that the Type I errors of the proposed approaches are slightly conservative and the detection rates are sensitive to the magnitude of the reduction in response time as a result of item preknowledge.

The results of the conditioning imposed in this study have the effect of reducing the operational number of examinees needed for estimating the effective response. It is suggested that ERT should be established only for those items with a sufficiently large number of examinees who were administered these items. Secondly, the conditioning imposed on the items included in computing X_c in Equation 11 has a similar effect of reducing the number of operational items. A possible way out is to find a suitable distribution G for a transformation $F(X_c) = v$ such that G is symmetric. Then, we can obtain a left-sided test by computing $\Pr(G(X_c) \leq x) = p$ without eliminating the other items.

Several actions could be taken when an examinee is suspected of item preknowledge: (1) instead of reporting one ability estimate for an examinee, several ability estimates could be reported on the basis of subtests that are in agreement with the model; (2) the item score pattern may be modified (for example, eliminate the unreached items at the end) and θ may be re-estimated; (3) the score could be withheld from reporting and the examinee may be retested; or (4) take no action if the error is small enough for the impact on the ability estimate to be marginal. This decision can be based on a comparison of the error introduced by the measurement disturbance to the standard error associated with each ability estimate. Which of these actions is taken very much depends on the context in which testing takes place. The usefulness of person-fit statistics thus also depends heavily on the application for which it is intended.

References

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 6*, 152–165.

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (ETS Technical Report No. 96-4). Princeton, NJ: Educational Testing Service.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*, 359-375.

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23*, 147–160.

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

Steinberg, J. (2002, August 8). Officials link foreign web sites to cheating on graduate admission exams. *The New York Times*, p. A18.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer Verlag.

Schnipke, D. L., & Scrams, D. J. (1997). *Representing response-time information in item banks* (LSAC Computerized Testing Report No. 97-09). Newtown, PA: Law School Admission Council.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57–75.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277–292.

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.

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index in multiple choice testing. *Journal of Educational Measurement, 39*, 115–132.

Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, *40*, 53–69.

Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 178–202). New York: Academic Press.

Veerkamp, W. J. J. (1996). *Statistical methods for computerized adaptive testing*. Unpublished doctoral dissertation, University of Twente, The Netherlands.

Van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas (Eds.), *New developments in computerized adaptive testing: Theory and practice* (pp. 201–219). Boston: Kluwer-Nijhoff Publishing.

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

van der Linden W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika*, *68*, 251–265.

van der Linden, W. J., Veldkamp, B. P., & Reese, L. M. (2000). An integer programming approach to item pool design. *Applied Psychological Measurement*, *24*, 139–150.

Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21*, 307–320.