Utilization of Response Time in Data Forensics of K-12 Computer-Based Assessment

XIN LUCY LIU

Data Recognition Corporation (DRC)

VINCENT PRIMOLI

Data Recognition Corporation (DRC)

CHRISTIE PLACKNER

Data Recognition Corporation (DRC)

Paper presented at the 2013 Conference on the Statistical Detection of Potential Test Fraud. Madison: WI. **Please address all questions to Xin Lucy Liu LLiu@DataRecognitionCorp.com.**

The authors thank Julie Korts from DRC for her valuable professional assistance in editing and formatting.

UTILIZATION OF RESPONSE TIME IN DATA FORENSICS OF K-12 COMPUTER-BASED ASSESSMENT

ABSTRACT

In the current study, the authors investigated the procedures and utility of using item response time to detect aberrant test behaviors on K-12 online state tests. When an unexpected short item response time produces an unexpected correct response, it may indicate that the examinee has some preknowledge of the item. Given the ability of a person (estimated by an IRT model) and the speediness of an item (predicted by a loglinear response time model), the expected response time (called effective response time, ERT) was estimated through an ERT data set selected by two criteria: the item was answered correctly and that the probability of item being answered correctly is large enough. For each item, the divergence of the observed response time from the expected ERT was tested by a chi-square statistic against the standard normal distribution. Furthermore, the item-level statistics were aggregated to subgroups of items classified by item difficulty and item type to examine whether there was differential impact of these item properties on the likelihood of aberrant responses. In addition, the item-level statistics were aggregated to student-level and school-level to identify suspicious examinees or schools.

The proposed procedures were applied to a real data set from a state assessment test and provided insightful findings regarding online test fraud detection for K-12 state assessment programs. Higher likelihood of aberrant responses seems to be more related to multiple-choice items rather than constructed-response items. In terms of item difficulty level, relatively hard items seem to be more associated with the tendency for cheating. Content examinations of the identified items reveal important implications to test development. The findings from the item

response time approach were cross-validated with those from the other statistics to detect aberrant behavior, such as wrong-to-right answer changes and counts of item visits, which indicates that the ERT approach is a promising method to distinguish items or examinees with aberrant responses from those with regular responses.

INTRODUCTION

In many states across the country, the testing mode in K-12 state assessment tests is gradually shifting from traditional paper-and-pencil testing to computerized testing. Due to the high stake nature of the state tests (e.g., its impact on school funding and teacher evaluations), data forensics analysis is necessary to ensure the integrity of test results. Compared to paper-and-pencil testing, online testing poses new challenges to test security. But at the same time, additional test information becomes available, e.g., item test time, number of visits, wrong-to-right answer change, etc., that may not be attainable from paper-pencil testing. Response times (RTs) have been an important source of information utilized in detecting aberrant test responses in online testing. Usually a typical pattern of RTs is expected for a particular item or a selected set of items. Unexpected RTs may be indicative of some specific aberrant behaviors. Typically, when a short item RT produces an unexpected correct response, it indicates that the examinee might have some preknowledge of this item. Preknowledge may be obtained from unauthorized acquisition and disclosure of high stake test materials (e.g., item teaching), which will undermine the accuracies of inferences from the test scores.

To our knowledge, most of the publications on the use of RTs for checking aberrances on online testing programs are based on testing the difference of the observed RTs from the predicted RTs (e.g., van der Linden & van Krimpen-Stoop, 2003; Meijer & Sotaridona, 2006;

van der Linden & Guo, 2008, and Meng, Li, & Steinkamp, 2011). These studies differ in the procedures and models adopted in the parameter estimation of the predicted RTs that the examinee needed to process the items to produce a response. Van der Linden and van Krimpen-Stoop (2003) adopted a Bayesian prediction of the RTs from the observed RTs and a normal prior on examinee slowness parameter. Meijer and Sotaridona (2006) estimated the expected RTs via an effective response time data selected for two criteria: 1) examinees' responses should be correct; and 2) the probability of answering an item correctly should be large enough. The item-level test statistics proposed by these authors were further aggregated to person levels in detecting suspicious examinees with aberrant test behaviors in the study by Meng, Li, and Steinkamp (2011). The three studies mentioned above used the same RT model that modeled the logarithm of response time via a linear composition of its person (slowness or speediness) and item (time intensity) parameters, as proposed in Thissen (1983) and van der Linden and van Kimpen-Stoop (2003). On the other hand, van der Linden and Guo (2008) adopted a new RT model proposed by van der Linden (2006) which has a parameter structure analogous to that of a 2PL logistic IRT model - a discrimination parameter that is modeled on the relation between item response time and person speediness from item time intensity. Based on the posterior predictive distribution of the RTs, the authors identified aberrant responses (i.e., preknowledge) if the probability of a predicted RT was lower than the observed RT. Similarly, the erratic RTs (e.g., memorization) are detected when the probability of the observed RTs exceed the predicted RTs.

The current study's main aim is to apply item RTs in detecting aberrant responses on K-12 online state assessments. Unlike computer adaptive tests in licensing fields, the K-12 state assessments are mostly computer-based. Students could answer the items in any order by using

navigation tools. They could omit items (i.e., see an item but not answer it and proceed to another), and they could go back to previously viewed items and change their answers. The tests were administered in this fashion to be as similar as possible to the paper-and-pencil mode of test delivery. The recorded response time for an item was the total time spent on the item during all attempts, as it was proposed by Schnipke & Scrams (1997). The recorded item visit frequency will be two or more for those who go back and review the item and one for those who visit the item, give an answer, and never come back to change it. The data on wrong–to-right answer change refer to the last change status of the item answer by those who visit the item two or more times. This variable is of binary values (1 for wrong-to-right change and 0 otherwise) and only recorded for multiple-choice (MC) items.

As discussed earlier, a few RT models were developed to incorporate the RTs in detecting aberrant responses. The lognormal model has found a good fit of its distributions to actual response time in many studies (Schnipke and Scrams, 1997; Schnipke & Scrams, 1999; Thissen, 1983; van der Linden, Scrams, & Schnipke, 1999). The Bayesian parameter estimation needs incorporate prior information on some parameters, for example, examinee slowness, which may not always be accessible in real-life settings. Therefore, we will adopt the loglinear model (van der Linden & van Krimpen-Stoop, 2003) and the classical approach of Effective Response Time (ERT) (Meijer and Sotaridona, 2006) in modeling and estimating the RTs.

The K-12 assessment tests in many states involve more item types other than multiple-choice type. The other commonly adopted type is Constructed Response (CR) items for which students need to produce an answer instead of choosing one from a few available options. To the best knowledge of the authors, few studies in the current literature have examined how certain properties (i.e., type, difficulty, etc) of an item may impact the tendency for aberrant behaviors.

As Ferrara (1997) pointed out, although cheating on the performance assessment (of which CR item is a kind) by test administrators could be somewhat more difficult than cheating on MC items using strategies of erasures or hints. Performance assessments tend to be more susceptible to disclosure than multiple choice tests because they tend to be more memorable, more like instructional tasks, and smaller in numbers, and thus it is more likely to incite or involve test administrators to violate test security. However, the susceptibility to cheating without detection for constructed responses seems substantially exceeds that for multiple choice tests. The cause for this unfortunate fact, according to the authors, is the lack of efficient methodology suited for this type of item. As constructed response items are widely adopted in K-12 state assessment as a way of measuring complex skills, the need to consider the test security issues with CR items is equally important as the need for the MC items. As more and more K-12 state testing is being made available on-line, test information on response time, visit frequency, and wrong-to-right answer change are produced for each individual item. The available item-level test information will provide a proper way to detect the susceptibility to cheating for different item types. Specifically, the purposes of this study are described in more details as follows:

1. The procedures of the ERT approach will be elaborated for detecting the aberrant responses at three levels of interest: item, student, and school.
2. The other two sources of information, wrong-to-right answer changes and item visit frequency, will be analyzed for two purposes:
    i. to cross validate the effectiveness of the ERT approach; and
    ii. to investigate their utility in data forensics.

3. The impact of different item types/difficulty levels on the likelihood of aberrancy will be examined using the ERT approach, and cross validated with the other two sources of information.

The methodologies of the ERT approach will be illustrated in the next sections, followed by the data descriptions, results, and discussions.

RESPONSE TIME MODEL

According to van der Linden & van Krimpen-Stoop (2003), a loglinear model was proposed to fit the RTs:

$$\ln T_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij} \quad (1)$$

with

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

where $\ln T_{ij}$ is the natural logarithm of the time taken by examinee *j* to respond to item *i*, δ_i is the parameter for the response time required by 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} a normally distributed residual or interaction term for item *i* and examinee *j* with the 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 1 can be estimated as follows:

$$\mu \equiv E_{ij}(\ln T_{ij}) \quad (3)$$

$$\delta_i \equiv E_j(\ln T_{ij}) - \mu \quad (4)$$

$$\tau_j \equiv E_i(\ln T_{ij}) - \mu \quad (5)$$

$$\sigma^2 \equiv E_{ij}(\ln T_{ij} - \delta_i - \tau_j)^2 \quad (6)$$

EFFECTIVE RESPONSE TIME

According to Meijer and Sotaridona (2006), the "effective response time" (ERT) is the time an individual examinee *j* with an ability level θ_j used to answer an item *i* correctly. To establish the ERT for each item *i* for each examinee *j*, we need to select a set of examinees for item *i* whose responses meet two requirements: 1) Item *i* should be answered correctly; and 2) Given an examinee's θ_j, the probability of item *i* being answered correctly should be large enough $P_i(\theta_j) > \gamma$ (for rational behind the two requirements, see the reference).

Given the ERT data, the ERT for each item *i* for each examinee *j* is modeled by regressing $\ln T_{ij}$ on θ_j and τ_j:

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

where the β's are regression coefficients, ε_j is an error term assumed to be normally distributed with mean 0 and variance σ_i^2. Thus, the expected RT is:

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

Observed RTs that are significantly different from expected can be used as evidence of item preknowledge. Then the observed RT of examinee *j* to item *i* is evaluated against the expected RT for that item by:

$$z_{ij} = \frac{\ln T_{ij} - \ln\hat{T}_{ij}}{\sigma_i}, \tag{9}$$

where σ_i^2 is the variance of the logarithm RT for item *i*:

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

We assume that the RTs are normally distributed in a log scale. It follows that z_{ij} is a standard normal distribution and z_{ij}^2 is Chi-Square distributed with one degree of freedom. The sum of the z_{ij}^2 across items taken by examinee *j* will hence follow a Chi-Square distribution with the degree of freedom equal to the number of items in the summation:

$$X_J = \sum_i^n z_{ij}^2 \sim \chi_{IJ}^2. \quad (11)$$

The quantity $Pr(X_j \geq x) = p$ will be compared to a significance level $\alpha = .05$ and $\alpha = .01$ respectively. The value of *p* that is less than α is indicative of significant aberrant responses.

THE DATA

Grade 4 Mathematics scores on 2012 state wide standardization test were available for 1,624 students from 38 schools and 27 districts. The median number of students per school was 41, with a range of 1-113.

Grade 4 Mathematics test includes 29 operational items, with 10 CR items and 19 MC items. Four of the CR items are in 3-point scale and the remaining 6 items are in 2-point scale. The raw scores in this sample range from 3 to 33 with the mean 19 and standard deviation 7.

There is no time limit for this test. So the test is not speeded. On average, examinees spent about 1 hour and 15 minutes on the whole test, and about 2.67 minutes per item. On average, CR items take more time to complete than MC items. Examinees spent about 3.6 minutes per CR item and about 2 minutes per MC item. The item time statistics are provided in Table 1. For the majority of the items (75.8%), examinees were able to complete an item within 2 minutes. On average, more difficult items take more time than easy items (see Table 2).

Examinees spent about 2 minutes per item on easy items with logits < 0 and about 3 minutes per item on more difficult items with logits ≥ 0.

[Insert Table 1 about here]

[Insert Table 2 about here]

CALIBRATION

The 19 MC items were calibrated with the 3PL logistics IRT model and the 10 CR items were calibrated using Generalized Partial Credit model. These two models were chosen due to their better fit of the data. According to the 3PL logistic model (Lord, 1980, chap. 5; van der Linden & Hambleton, 1997), the probability of a correct response on item *i* by person *j* is given by:

$$P\{U_{ij}=1\} \equiv p_i(\theta_j;\ a_i, b_i, c_i) \equiv c_i + (1-c_i)\frac{exp[a_i(\theta_j - b_i)]}{1+exp[a_i(\theta_j - b_i)]}, \quad (12)$$

where U_{ij} is a response indicator of examinee *j* to item *i* (1 if correct and 0 if incorrect), $\theta_j \epsilon R$ is the ability of test taker *j*, and $a_i \epsilon (0,\infty)$, $b_i \epsilon R$, and $c_i \epsilon [0,1)$ are the slope, location, and guessing parameters for item *i*, respectively.

According to the Generalized Partial Credit Model (Muraki, 1992), for item *i* scored in categories 0, 1, …, *k*, where *k* is the highest score category for the item, the probability of selecting the *k*th category from m_i possible categories of item *i* is given by:

$$P_{ik}(\theta) = \frac{exp[\sum_{v=0}^{k} Da_i(\theta - b_i + d_{i,v})]}{\sum_{n=0}^{m_i} exp[\sum_{v=0}^{n} Da_i(\theta - b_i + d_{i,v})]}, \quad (13)$$

where D is a scaling constant set to 1.7 to approximate the normal ogive model, α_i is a slope parameter, b_i is an item location parameter, and $d_{i,v}$ is a category parameter.

ABERRANT RESPONSE TIME

For each MC item *i*, the ERT data of examinees *j* were selected such that 1) item *i* is answered correctly, i.e., $U_{ij} = 1$; and 2) given examinee's θ_j, the probability of correct response is greater than the item guessing parameter estimated from Equation 12: $p_i(\theta_j) > c_i$. In contrast, for each CR item, the ERT data were selected based on only one criterion, that is, item *i* is answered with the maximum score points (e.g, 2 for 0-2 scored CR items, and 3 for 0-3 scored CR items). It is assumed that it is not possible for an examinee to guess a response for CR items and hence the time an examinee spends in successfully completing a CR item with full credits would reflect the effective response time needed for that item.

Given the ERT data, examinee's slowness parameter τ_j is estimated from Equations 3 and 5, and the expected RT $\ln\hat{T}_{ij}$ and the RT variance σ_i^2 were obtained from Equations 7, 8, and 10. The Chi-Square statistics z_{ij} that test whether the observed RT ($\ln T_{ij}$) is significantly different from the expected RT ($\ln\hat{T}_{ij}$) was computed for each item for each examinee using Equation 9. A total Chi-Square statistic χ_{IJ}^2 in Equation 11 was computed for each examinee by summing z_{ij}^2 over items for which the observed ($\ln T_{ij}$) was significantly different from the expected ($\ln\hat{T}_{ij}$), and then tested against the two criteria: $\alpha = .05$ and $\alpha = .01$ respectively. The degree of freedom for χ_{IJ}^2 is the number of items in the summation.

CROSS VALIDATION

As mentioned earlier, in addition to response time per item, additional sources of information are collected on computerized testing: the total number of item visits and the last answer change from wrong to right. On paper-pencil state assessment tests, schools' erasure rates

(number of changes from wrong-to-right over items and students) have been found to be related to school AYP status. Primoli (2012) pointed out that schools' probability of erratic erasure rates increases as their AYP failure severity increases (note that 'erratic' here means 'significantly higher than state average'). Plackner and Primoli (2012) further showed a moderate to high correlation between the aberrant erasure rates with other indicators for aberrant test behaviors. Given these findings in paper-and-pencil data forensics studies, let us hypothesize that on computerized testing, aberrant test responses are related to higher rate of answer changes from wrong to right, as such, to more number of item visits as well.

To cross validate the ERT approach in detecting aberrant test responses, the average number of wrong-to-right answer changes and the average number of item visits will be computed for each examinee over the items he/she takes. Then the two statistics will be compared between the group flagged by the total Chi-Square statistic χ_{IJ}^2 and the group not flagged. The result on the consistencies among the ERT approach and the other two sources of information will provide evidence for the validity and utility of these statistics in on-line forensics studies.

RESULTS

The summary statistics for the IRT item parameters are presented in Table 3 (note that guessing parameters are not applicable for CR items). On average, CR items are more difficult than MC items, though the slope parameters are similar in the magnitude and the distribution. The guessing parameters for MC items range from .134 to .403, and are used as one of the criterion for selecting ERT data for MC items.

[Insert Table 3 about here]

The size of each ERT data including examinees with both shorter $\left(\ln T_{ij} - \ln \hat{T}_{ij}\right)$<0 and longer $\left(\ln T_{ij} - \ln \hat{T}_{ij}\right) \geq$0 RT for each item ranges from 387 to 1313 examinees with a mean of 978 and a standard deviation of 252. The size of each ERT data that includes only the examinees with shorter RT for each item ranges from 216 to 707 examinees with a mean of 515 and a standard deviation of 135. These results indicate a sufficiently large number of examinees used to estimate the expected ERT for each item.

ITEM-LEVEL

Using each ERT data, we computed the Chi-Square statistics z_{ij}^2 for each item for each examinee and tested it against the standard normal distribution. To examine the impact of item type or difficulty on the likelihood of aberrant response, let's define the percentage of examinees flagged for significantly different RTs as item aberrancy rate for each item. The summaries of the aberrancy rate by item types and difficulty levels are reported in Table 4. On average, the MC items have higher percentage of aberrancy rate than the CR item at both levels of $\alpha = .05$ and $\alpha = .01$. The more difficult items (logits≥0) have a slightly higher aberrancy rate than the easy items (logits <0) at both alpha levels. The number of wrong-to-right answer changes and item visits are reported for different item types and difficulty levels in Table 5. Note that the data of wrong-to-right answer changes are only available for MC items. CR items or more difficult items are generally visited more frequently than MC items or easy items. More difficult items are on average associated with more wrong-to-right answer changes.

[Insert Table 4 about here]

[Insert Table 5 about here]

STUDENT-LEVEL

According to the total Chi-Square statistic χ^2_{IJ}, out of the 1624 examinees, the number of examinees identified using $\alpha = .05$ is 95 examinees (about 5.8%) and 47 (about 2.9%) using $\alpha = .01$. The number of wrong-to-right answer changes and the number of item visits were compared for flagged groups versus non-flagged groups in Table 6. On average, the number of item visits by flagged groups is greater than that by non-flagged groups at both the levels of $\alpha = .05$ and $\alpha = .01$. The flagged groups have higher frequencies of wrong-to-right answer changes than non-flagged groups at both alpha levels.

[Insert Table 6 about here]

Figure 1 is a plot of observed RTs (LnT for $\ln T_{ij}$ shown in red squares) over expected RTs (ElnT for $\ln \hat{T}_{ij}$ presented by blue diamonds) on correctly answered items by an examinee identified for his/her aberrant RT patterns $\chi^2(12) = 67.23, p < .01$. This examinee's ability parameter is -.32 logit. The x-axis is for the difficulties of the taken items ranging from -1.23 to .78. The y-axis represents the item RTs in natural logarithm. As shown in the figure, of the 12 items answered correctly, four were responded with unexpectedly shorter time over the difficulty range of .24 to .49. These four items are more difficult items relative to the person ability, but answered correctly in a much shorter time than expected, and thus resulted in a significant χ^2_{IJ}.

[Insert Figure 1 about here]

SCHOOL-LEVEL

The percentage of examinees flagged per school (called school aberrancy rate, for convenience) is computed for $\alpha = .05$ and $\alpha = .01$. Out of the 38 schools, 17 schools (N=982)

have aberrancy rate equal to or above 5%; the remaining 21 schools (N=642) less than 5% for $\alpha = .05$. For $\alpha = .01$, 8 schools (N=400) with equal to or above 5% examinees are flagged; and the remaining 30 schools (N=1224) with less than 5% examinees identified. The summary statistics for number of wrong-to-right answer changes and the number of visits for these two groups of schools are reported in Table 7. On average, schools with equal to or greater than 5% students identified produced higher number of wrong-to-right answer changes or item visits than schools with less than 5% identified at both $\alpha = .05$ and $\alpha = .01$.

[Insert Table 7 about here]

DISCUSSIONS

In this study, we investigated the procedures and the utility of the ERT approach by using RTs in detecting online test aberrant behaviors in a K-12 state assessment Grade 4 mathematics test. The ERT approach was also applied to investigate the question whether specific item types (e.g., MC or CR) or varying item difficulties are more prone to aberrant test response.

The findings indicated a higher likelihood of the MC items being answered correctly in an unexpected RT, compared to the CR items. This is probably because this type of item format might be easier to cue than CR items. Relatively more difficult items of either item type are more associated with aberrant response time. This makes sense as there is basically no need to cheat on very easy items. The CR items were visited more often than the MC items. This indicates that the frequencies of item visits are probably more related to the difficulty of the items rather than the tendency for cheating. Thus, consistency between the RT results and the item visits do not hold true at the item level for different item types. However, the consistency

between the ERT flag results and the wrong-to-right answer changes / item visits, generally holds true in identifying aberrant responses at a student or school level. It indicates that the ERT approach is promising in distinguishing examinees with aberrant test behaviors from those with regular test behaviors on online testing.

In taking a closer look, many of the items with an ERT flag seem to have a more distinct theme than items not identified and therefore may be more subject to memorization. For example, there is a mathematics item on counting red roses. An easy cuing might be that "if you see the red rose item, the answer is B". Such content findings might have practical implication to item development. Perhaps during the item development stage it is a good idea to avoid creating items that are easy to memorize or communicate, or to avoid unusual phrases or theme-like stems especially for MC items.

Future research may answer some remaining issues. First, the available sources of information collected from online testing may not be equally effective for forensic purposes. This study indicated that the number of item visits may not be a good indicator as effective as wrong-to-right answer changes. Second, there are two subtypes of CR items in this data set: 1) SA (Short Answer) scored in 3 points, and 2) CP (Completion) scored in 2 points. They are not analyzed separately due to the small size (4 items for SA and 6 items for CP). Future research might need to look into the effect of item subtypes on forensics studies. Third, the findings in the current study are based on a single data set from a Grade 4 state Mathematics test. Therefore, the findings might not be generalizable to other populations, grades, or contents areas. We suggest that the similar procedures should be applied to other grades and subject areas to investigate the effectiveness of the ERT approach and the relation of the item type/difficulty to test aberrancy. Finally, in this study we assume that if a CR item was answered with a maximum score point,

there is no cheating. This assumption might be true for low-performing students because even with item teaching, they might still be unable to achieve a maximum score point on a CR item. However, for medium- or high-performing students, it is possible that they might gain the full credit from item teaching. Therefore, for future forensic studies of CR items, some potential factors (e.g., subtypes or student ability, etc) need to be considered in order to obtain a thorough understanding of the aberrant item response on CR items.

REFERENCES

Ferrara (1997). *Test security for high stakes performance assessments: Consideration of ethics, validity, and data integrity*. Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, Chicago, IL.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Meijer, R. R. & Sotaridona, L. S. (2006). *Detection of Advance Item Knowledge Using Response Times in Computer Adaptive Testing* (LSAC Computerized Testing Report No. 03-03). Newtown, PA: Law School Admission Council.

Meng, H., Li, X., & Steinkamp, S. (2011). *Detecting Aberrant Responses in Computer-Based Testing*. Paper presented at the 2011 Annual Meeting of the National Council on Measurement in Education, New Orleans.

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

Plackner, C., & Primoli, V. (2012). *Data forensics: A compare and contrast analysis of multiple methods*. Paper presented at the Conference on Statistical Detection of Potential Test Fraud, Lawrence, KS.

Primoli, V. (2012). *AYP consequences and erasure behavior*. Paper presented at the Conference on Statistical Detection of Potential Test Fraud, Lawrence, KS.

Schinipke, D. L., & Scrams, D. J. (1997). Modeling Item Response Times With a Two-State Mixture Model: A New Method of Measuring Speededness. *Journal of Educational Measurement*, *34*(*3*), 213-232.

Schinipke, D. L., & Scrams, D. J. (1999). *Representing response-time information in item banks* (LSAC Computerized Testing Report No. 97-09). Newtown PA: Law School Admission Council.

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.

van der Linden, W.J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181-204.

van der Linden, W.J. (2009). A Bivariate Lognormal Response-Time Model for the Detection of Collusion Between Test Takers. *Journal of Educational and Behavioral Statistics*, *34*(3), 378-394.

van der Linden, W.J., & Guo, F. (2008). Bayesian Procedures for Identifying Aberrant Response-Time Patterns in Adaptive Testing. *Psychometrika*, *73*(*3*), 365-384.

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., Scrams, D. J.,. & Schnipke. D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195-210.

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*(*2*), 251-265.

Table 1. *Item Time Statistics.*

Item Time Category	MC	CR	Total	%	Cumulative %
=< 1 minute	5	0	5	17.2%	17.2%
1 -2 minutes	12	5	17	58.6%	75.8%
>= 3 minutes	2	5	7	25.2%	100%

Table 2. *Item Time Statistics by difficulty*.

	MC		CR	
	Easy	Hard	Easy	Hard
# of Items	11	8	4	6
Avg Time	1.88	2.96	1.95	2.87

Table 3. *Item Parameter Summary Statistics.*

	Statistics	Location	Slope	Guessing
CR (Items=10)	Min	-1.327	.481	
	Max	.771	1.263	
	Mean	.068	.818	
	SD	.680	.270	
MC (Items=19)	Min	-1.298	.395	.134
	Max	.654	1.348	.403
	Mean	-.386	.808	.209
	SD	.650	.267	.065

Table 4. *Summary Statistics for Aberrancy Rate by Item Type and Difficulty*.

		Type		Difficulty	
		MC	CR	Easy	Difficult
$\alpha = .05$	Min	1.314	.926	1.314	.926
	Max	5.038	4.039	4.115	5.038
	Mean	2.979	2.654	2.539	2.898
	SD	.942	1.237	.895	1.215
$\alpha = .01$	Min	.000	.000	.000	.000
	Max	2.519	1.292	2.147	2.519
	Mean	.814	.509	.745	.870
	SD	.645	.481	.530	.691

Table 5. *Summary Statistics for Wrong-to-right Changes and Number of Visits by Item Type and Difficulty*

		Type		Difficulty	
		MC	CR	Easy	Difficult
W-to-R Count	Min	0		0	0
	Max	23		23	12
	Mean	.08		.06	.09
	SD	1.119		1.167	.351
# of visits	Min	1	1	1	1
	Max	17	14	13	17
	Mean	2.42	2.59	2.40	2.59
	SD	1.989	2.053	1.951	2.097

Table 6. *Summary Statistics for Number of Wrong-to-Right changes and Number of Visits for Flagged and Non-Flagged Groups of Examinees.*

		Flagged		Non-flagged	
	Statistics	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
W-to-R Count	Min	.00	.00	.00	.00
	Max	4.66	1.45	14.59	14.59
	Mean	.15	.11	.11	.08
	SD	.51	.22	.88	.87
# of Visits	Min	1.07	1.17	1.00	1.00
	Max	10.24	10.24	11.76	11.76
	Mean	3.12	3.44	2.71	2.71
	SD	2.25	2.38	1.81	1.82

Figure 1. Flagged Example Examinee (person theta = -.32)

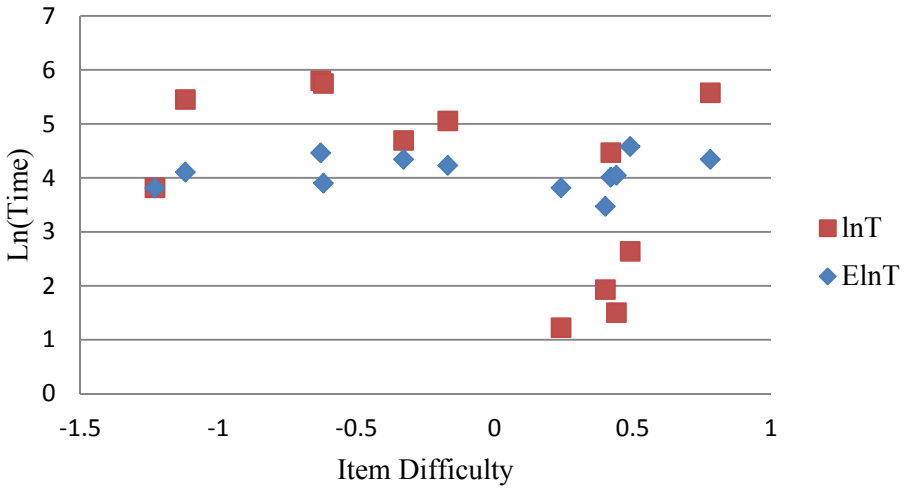

Table 7. *Summary Statistics for Number of Wrong-to-Right changes and Number of Visits for Two Different School Aberrancy Rates.*

		School Aberrancy ≥ 5%		School Aberrancy < 5%	
	Statistics	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
W-to-R Count	Min	.000	.000	.000	.000
	Max	14.586	14.586	6.862	6.862
	Mean	.135	.202	.075	.082
	SD	1.070	1.622	.309	.327
# of Visits	Min	1.000	1.000	1.000	1.000
	Max	15.759	14.276	12.655	13.759
	Mean	2.819	2.741	2.507	2.753
	SD	1.990	2.076	1.620	1.777