

An Investigation of Rolling Person Fit in Identifying Examinees Who Abandon Test Effort

Steven L. Wise

Northwest Evaluation Association

Christine E. DeMars

James Madison University

Paper presented at the 2008 meeting of the International Psychometric Society, Durham, New Hampshire, USA.

Abstract

When the perceived benefit of a test (to self or others) is low relative to the perceived effort required, some examinees will abandon test effort and respond randomly. Three person fit indices were modified to identify the item location where an examinee abandoned test effort: ℓ_z , marginal likelihood ratio (random/solution behavior), and conditional (conditional on θ) likelihood ratio. The modified indices were termed rolling person fit because at each test item the index was calculated based only on responses to the 10 most recently administered items. When an examinee responded randomly to most of the recent items, the conditional likelihood ratio and ℓ_z had good power if the examinee had moderate or high proficiency. The marginal likelihood ratio had better power for examinees with lower proficiency but also higher Type I errors for low-proficiency examinees. With real data, examinees who exhibited rapid-guessing behavior to a string of items also had poor rolling person fit for those items. This validates the interpretation of rapid guessing as random responding. However, with the low base rate of rapid guessing in the data, most of the response strings identified as poor fitting were not rapid guesses.

An Investigation of Rolling Person Fit in Identifying Examinees Who Abandon Test Effort

When an achievement test is used to measure an examinee's level of proficiency, it is implicitly assumed that the examinee devotes sufficient effort to the test items to ensure that the resulting test score accurately reflects his or her actual level of proficiency. Without adequate effort, test performance is likely to suffer, resulting in the test score underestimating the examinee's actual level of proficiency. Low examinee effort can lead to a negatively biased proficiency estimate, and thereby constitutes a source of construct-irrelevant variance that threatens the validity of test scores (Haladyna & Downing, 2004).

This validity threat is most likely to pose a problem whenever there are little or no test performance-based consequences to examinees. We will term these *low-stakes* tests, in reference to the examinee's perspective (even when there are substantial test-performance-based consequences for those giving the test). If examinees are not motivated to give their best effort to their test items, then their demonstrated proficiency levels (i.e., their score-based proficiency estimates) are apt to be lower than their actual proficiency levels. Numerous studies have shown that examinee test-taking motivation can have a substantial influence on test performance. Wise and DeMars (2005) provided a synthesis of much of this research, finding that motivated examinees tend to outperform their unmotivated counterparts, on average, by over a half standard deviation.

Test-taking motivation, however, is not an all-or-none proposition when low-stakes tests are used. Some examinees will try hard initially, but at some point they may choose to abandon their effort. Wise and Smith (in press) proposed a model of test-taking effort in which an examinee's capacity to expend effort on a test is a function of both the consequences associated with the test as well as a variety of internal factors, including academic citizenship,

competitiveness, ego satisfaction, and a desire to please teachers, parents, and others. Under the Wise-Smith model, test givers implicitly rely on these internal factors to motivate examinees on low-stakes tests. Unfortunately, this reliance by test givers on internal motivational factors can result in “test abandonment” being exhibited by examinees for whom internal factors are not sufficiently strong.

An example can illustrate how test abandonment might occur. Imagine that you have asked a university sophomore to take a low-stakes science achievement test that is to be used for academic program assessment. The student, being a good academic citizen, agrees to take your test. If you then administer a test consisting of 20 multiple-choice items, the chances are good that she will give good effort to this brief test. In contrast, if you had administered a test consisting of 200 multiple-choice items, the student tries hard for a while, but at some point she gives up and stops trying. According to the Wise-Smith model of test-taking motivation, the student would have enough “effort capacity” to effortfully complete the shorter test but not enough to complete the longer one. In essence, by administering the longer test we had asked for more effort than the student was willing to give, and test abandonment occurred. Consequently, a proficiency estimate based on her responses would likely underestimate her actual proficiency level.

The identification of test abandonment requires that the test giver have the ability to measure examinee effort. Until recently, effort was typically measured using a post-test self-report instrument. These types of measures, however, have two disadvantages. First, self-report measures are limited to measuring an examinee’s perception of the general level of effort exhibited across the testing session. Second, self-report measures are vulnerable to various types of biases. For example, if an examinee tended to attribute failure to lack of effort, and he had

perceived that he had not performed well on the test he had just taken, then he might understate his level of test-taking effort, to rationalize his perceived poor test performance.

Wise and Kong (2005) introduced a new measure of examinee effort based on item response time (which requires a computer-based test [CBT]). They extended the research of Schnipke (1995; 1996; Schnipke & Scrams, 2002), who studied instances of examinees rapidly entering answers during speeded, high-stakes tests. As time is running out, some examinees will switch from trying to work out the answer to items (termed *solution behavior*) to rapidly answering all of the remaining items in the hope of guessing some correct answers (termed *rapid-guessing behavior*). Each item response can be classified as one of these two behaviors by comparing the response time to a pre-determined time threshold established for that item. Schnipke and Scrams (2002) provide a good overview of the research on rapid-guessing behavior in high-stakes tests.

Wise and Kong (2005) studied the response time data from unspeeded, low-stakes CBTs, and found that rapid guessing can occur throughout a testing session, and not just toward the end, as is typically observed with high-stakes CBTs. They argued that in low-stakes situations, rapid guesses represent non-effortful behaviors by unmotivated examinees. Wise and Kong developed a measure of examinee effort, termed *response time effort (RTE)*, which represents one minus the proportion of test items for which an examinee exhibited rapid-guessing behavior. In other words, RTE is the proportion of items to which the examinee applied solution behavior. Relative to self-report measures, the identification of rapid-guessing behavior has the advantage of being based on observed examinee behavior (and therefore is not vulnerable to many of the potential biases associated with self reports). More important to the current study, however, is the additional advantage that rapid-guessing behavior can be evaluated at the item level. This

permits examinee effort to be evaluated at different points in the test. For example, Wise and Kong showed a graph of item response times for an examinee who exhibited solution behavior for the first two-thirds of the test, and then switched to rapid-guessing behavior for the remainder of the items. In effect, the graph was consistent with an examinee exhibiting test abandonment.

Research on rapid guessing on low-stakes tests has yielded several important findings. These include (a) examinees can vary considerably in RTE, (b) an examinee's effort can change during a test, (c) RTE tends to be substantially correlated with test performance, and RTE tends to be uncorrelated with external measures of academic ability, such as SAT scores (Wise & Kong, 2005). In addition, the presence of rapid guesses can have a pronounced effect on the psychometric properties of test data. One effect is a decrease in test score validity (DeMars, 2007; Kong, 2007; Kong, Wise, Harmes, & Yang, 2006; Wise, Bhola, & Yang, 2006; Wise & DeMars, 2006; Wise & Kong, 2005). Another, possibly less obvious effect of rapid guessing is that the internal consistency of a set of scores can be spuriously inflated (Kong, 2007; Wise, 2006; Wise & DeMars, 2006, in press).

An important additional finding regarding rapid guesses is that they tend to be accurate about as often as random responses, regardless of the difficulty levels of the items to which they have been given (Wise & Kong, 2005). This suggests that person fit indices, which will generally identify random responses as aberrant (i.e., not fitting the measurement model being used), could also be used to identify test abandonment. Thus, test abandonment should be characterized by rapid responses that exhibit poor person fit.

The purpose of this study was to investigate the degree to which person fit can be used to identify examinees who abandon effort during a test. Even though item response time has been shown to be useful in identifying test abandonment, there are two reasons why it would be useful

to assess the usefulness of person fit. First, the collection of item response time is feasible only when CBTs is used, which restricts its practical use in identifying test abandonment. Person fit, in contrast, can be computed for both CBT and non-CBT data. Second, it is possible that test abandonment can occur, but rapid-guessing behavior is not exhibited. That is, there may be instances of examinees responding randomly but slowly to test items. Hence, person fit may be found to be a more comprehensive indicator of test abandonment than response time.

Person Fit Indices

Likelihood Ratio. When a model form can be specified for a particular type of aberrant responding, the likelihood that an observed response string follows the aberrant model can be compared to the likelihood that it follows the normal model. In this study, the model for aberrant response is the random-response behavior model and the normal model is the model for solution behavior. A ratio less than one indicates the responses were more likely under solution behavior, and a ratio greater than one indicates the responses were more likely under random-response behavior. Levine and Drasgow (1984) used a statistical test based on the ratio

$$L_A(\mathbf{x})/L_N(\mathbf{x}), \quad (1)$$

where $L_A(\mathbf{x})$ is the likelihood of response pattern \mathbf{x} under the model for aberrance and $L_N(\mathbf{x})$ is the likelihood under the normal model. For this study, the normal model (solution behavior) is the 3PL IRT model. The aberrant model is random response, such that the probability of correct response is 1/the number of response alternatives. Under this random response model, the probability of correct response does not depend on θ and $L_A(\mathbf{x})$ is simply the product of the probabilities of the observed responses under random-response behavior.¹ However, the

¹ This is much simpler than the aberrant models used in other work where the number of aberrant responses, but not the specific items responded to randomly, was specified for each examinee (Drasgow, Levine, & McLaughlin, 1987;

probability of correct response under solution behavior depends on θ and, to find the overall likelihood of a response pattern, the likelihood function must be integrated over the θ distribution to obtain $L_N(\mathbf{x})$, the marginal likelihood.

$$L_N(\mathbf{x}) = \int P(\mathbf{x} | \theta) f(\theta) d(\theta) \quad (2),$$

where $f(\theta)$ is the normal distribution or another appropriate distribution. Data can be simulated under the normal model to determine the appropriate cut-off value for a given Type I error rate, as illustrated in Drasgow, Levine, & Zickar (1996).

For this study, the likelihood conditional on θ (conditional likelihood ratio) replaced the marginal likelihood. In other words, $L_N(\mathbf{x}|\theta)$ replaced $L_N(\mathbf{x})$: the likelihood was not integrated across the population density $f(\theta)$. The conditional likelihood takes into account that, for very low θ , random responses are equally likely under both rapid-guessing and solution behavior. As Molenaar and Hoijtink (1990, p. 79) noted, a particular response pattern, in this case random responding, may be “very improbable for a randomly sampled respondent, but it is still the most probable answer for a respondent with very low ability”. Such patterns would be flagged as aberrant using the marginal likelihood ratio but not when using the conditional likelihood ratio. Reciprocally, the conditional likelihood takes into account that random responses are particularly unlikely for very high θ . Thus, the conditional ratio will have fewer Type I errors than the marginal ratio for low θ , and greater power than the marginal ratio for high θ . However, use of the conditional ratio with real data assumes that all examinees use a reasonable degree of solution behavior for at least the initial items used to estimate θ ; the first 20 items were used for

Levine & Drasgow, 1984; 1988). In the current study, the aberrant model specifies each of the previous 10 items as random.

this study. If an examinee is very unmotivated at the beginning of the test, the θ estimate for that examinee will be biased low and rapid-guessing behavior later in the test will not look aberrant. The marginal ratio would have higher power to detect this examinee as problematic. Further, estimation of θ introduces some additional error into the likelihood ratio.

The person-fit index ℓ_z . One index proposed by Levine and Rubin (1979) to detect unlikely response patterns was ℓ_0 , the log likelihood function of an examinee's response pattern. The distribution of ℓ_0 is not constant across θ , so Drasgow, Levine, and Williams (1985) standardized it by subtracting the expected value and dividing by the standard deviation. They labeled this index Z_3 and it has also become known as ℓ_z . They showed that it had a sampling distribution that was approximately, though not precisely, standard normal.

When θ is known, ℓ_z has a null distribution with a mean near 0 and a standard deviation near 1 (Nering, 1995; van Krimpen-Stoop & Meijer, 1999), except for θ values located where there is almost no test information (Reise, 1995). However, these same studies have shown that ℓ_z is negatively skewed, so that the Type I error rate is greater than would be expected in a normal distribution. When θ is estimated from the data, the null distribution of ℓ_z tends to have a mean > 0 and a standard deviation < 1 , but continues to be negatively skewed (Nering, 1995; Reise, 1995; van Krimpen-Stoop & Meijer, 1999). Depending on the balance of these factors, the test can be conservative or liberal. The mean and standard deviation become closer to 0 and 1, respectively, as the test length increases (van Krimpen-Stoop & Meijer, 1999), which is likely why Drasgow et al. (1985) found a distribution close to standard normal using 85 items.

Drasgow, Levine, and McLaughlin (1987) and Drasgow et al. (1985) explained that it would be difficult to detect random response patterns using ℓ_z for low proficiency examinees

because these responses would have little effect on the likelihood (the same issue noted for the likelihood ratio index). For example, with 30% random responses and a Type I error rate of .01, ℓ_z had a power of .95 for examinees with scores above the 93rd percentile, but only .35 for examinees between the 49th and 64th percentiles (Drasgow et al., 1987). They did not explore fit indices for random responses among lower proficiency examinees because the power was expected to be low.

Person Fit as a Rolling Index. A potential drawback to the use of person fit is that it typically is not calculated until the end of the test. This does not reveal *where* in the test the examinee abandoned effort. Also, at this point it is too late to attempt intervention to encourage the examinee to make more effort. For example, Wise et al. (2006) showed that when warning messages were displayed to examinees who had begun to exhibit rapid-guessing behavior, subsequent rapid guessing was reduced. If these warning messages were to be based on person-fit, person-fit would need to be estimated at points throughout the test. One option would be to calculate fit after every item, based on all of the items up to that point in the test. This procedure may take a while to detect that an examinee has abandoned test effort, particularly if abandonment occurs well after the test has begun. Stated another way, it is unclear how sensitive a person fit index would be when the response pattern contains both responses that fit the measurement model and some that do not. For example, suppose that an examinee gave good effort to the first 30 test items and then abandoned effort for the remainder of the test. Because the response pattern contained 30 responses that ostensibly fit the model, it would probably take quite a few random responses before misfit would be identified. Thus, whenever abandonment occurred in the middle of a test, the sensitivity of the person fit index would be of concern.

To address this problem, a *rolling fit index* was used, in which person fit would be calculated only over the k most recently administered items. Thus, if rolling person fit was calculated over the most recent 10 items in the example described above, after the 40th item had been administered, fit would be calculated only for the 10 most recent responses (items 31-40) and would not consider the initial 30 items that fit the model. In this fashion, it was presumed that the rolling fit approach would more quickly identify both that an examinee had abandoned test effort, and the approximate point during the test administration that test abandonment had occurred.

Purpose

In Study 1, the power of three rolling fit indices to detect random responding were compared: marginal likelihood ratio, conditional likelihood ratio, and ℓ_z . The marginal Type I error rates were kept constant across the procedures, and their power and Type I error rates conditional on θ were assessed. In Study 2, the relationship between rapid-guessing behavior and the detection of random responding was explored using real data from a CBT. Because rapid-guessing behavior is essentially random, it was hypothesized that periods of rapid guessing would be detected as random responding using the rolling fit measure.

Study 1

Method

Data Simulation. Three thousand test forms were simulated with 60 items each. Different items were randomly selected for each test form because preliminary work showed that the distributions of the fit indices differed somewhat depending on the specific item parameters; using 3000 different sets of parameters randomized this effect throughout the test. For each test form, simulees' θ s were uniformly distributed from -3 to 3 at intervals of .5. These same values

of θ were used for each test form so that distributions of each index could be estimated at each θ value. When results were reported as averages across θ levels, θ 's were weighted based on the standard normal density. Dichotomous responses (incorrect/ correct) were simulated under the solution-behavior model. Then, for the aberrant conditions, each record was copied and responses were randomly replaced beginning with item 11, 21, 31, 41, or 51.

Fit indices based on the previous 10 items were calculated beginning at item 10. Thus, each simulee had 51 values for each of the indices. Indices were calculated once conditional on the true θ and again conditional on θ estimated from the first 20 items. Obviously, the true θ would never be known with real data, but this condition was included to explore the properties of the fit indices under ideal conditions. EAP estimation, with a standard normal prior, was used for the θ estimates. The true item parameters were used in the calculations—this represents a context where the item parameters have been estimated previously in a large sample of motivated examinees using solution behavior. The indices studied were: Likelihood ratio conditional on true θ , likelihood ratio conditional on estimated θ , ℓ_z conditional on true θ , ℓ_z conditional on estimated θ , and the marginal likelihood ratio.

Flagging Criteria. For ℓ_z , examinees were flagged as misfitting when the index was < -2.58 , a nominal error rate of .005. However, because ℓ_z did not have a standard normal distribution, even when conditional on true θ , the empirical error rate in the null distribution was higher. After the mean Type I error rate was calculated for ℓ_z , the point was found in the weighted null distribution of the conditional likelihood ratio or the marginal likelihood ratio that would yield the same Type I error rate. Thus, the power of the indices could be compared with the same Type I error rate.

Results

True Theta. The true θ condition was included to explore the properties of the rolling fit indices under ideal conditions of known θ . Under the null assumption of solution behavior, the distribution of ℓ_z had a mean of 0 and standard deviation of 1, but was skewed to produce a heavier lower tail than would be seen in a normal distribution, as in other studies (Reise, 1995). Thus, using a cut-off value of <2.58 yielded an empirical Type I error rate of .017 instead of .005. To produce the same Type I error rate for the conditional likelihood ratio index, a cut-off value of 4.299 was selected.

In Figure 1, the detection rate for each index is plotted as a function of the number of random responses in the previous 10 items. Zero indicates only solution behavior, and ten indicates only random behavior. As would be expected, the detection rate increased with the number of random responses. The index ℓ_z had higher power when 5 or fewer of the 10 items were random responses, but the increase in power leveled off and the conditional likelihood ratio had higher power for more random responses. Also, ℓ_z detected almost 60% of examinees who responded randomly to the entire string of 10 items, and the conditional likelihood ratio detected 74%. The conditional likelihood ratio probably had lower power when fewer responses were random because the aberrant model used in the numerator was the likelihood of responding randomly to all 10 items (in contrast to Drasgow et al.'s (1987) aberrant model for a subset of random responses).

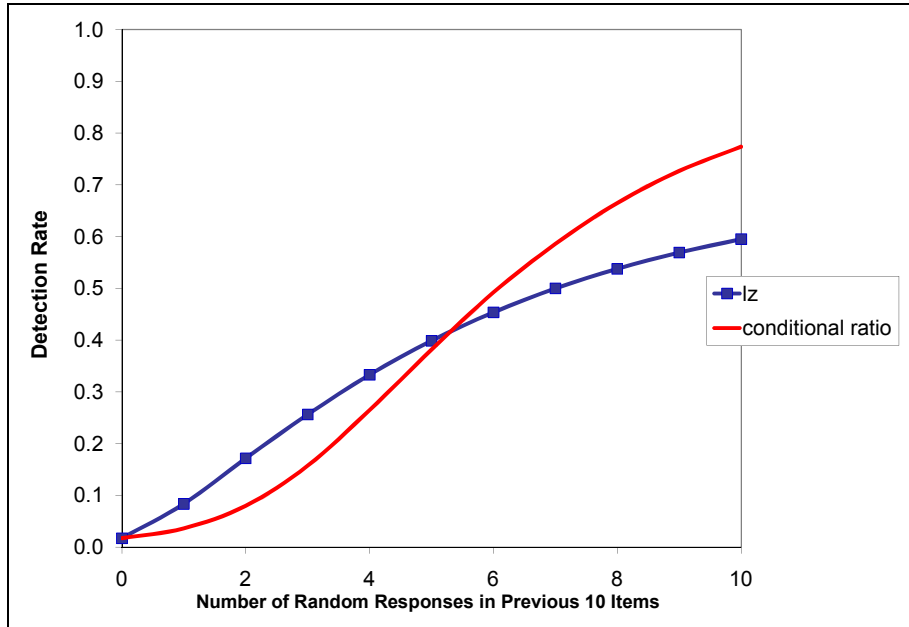


Figure 1: *Detection rates, based on true θ , as the number of random responses increases.*

Neither index was able to detect 100% of examinees who responded randomly. Random responses were much more difficult to detect for low-proficiency examinees because random response patterns tend to resemble the response patterns of low-proficiency examinees when they use solution behavior. Figure 2 illustrates the detection rate by θ for examinees who randomly responded to all 10 of the previous 10 items. The Type I error rate is also shown by θ . The conditional likelihood ratio virtually never detected very low-proficiency examinees ($\theta < -2$) either erroneously or correctly. The Type I error rate was slightly higher for $-2 < \theta < 0$. Random responses were less likely to occur by chance in this range, so when they did occur in the null condition they were falsely flagged. The Type I error rate for ℓ_z was more stable but increased slightly with θ . For $\theta > .5$, power was very high for both indices.

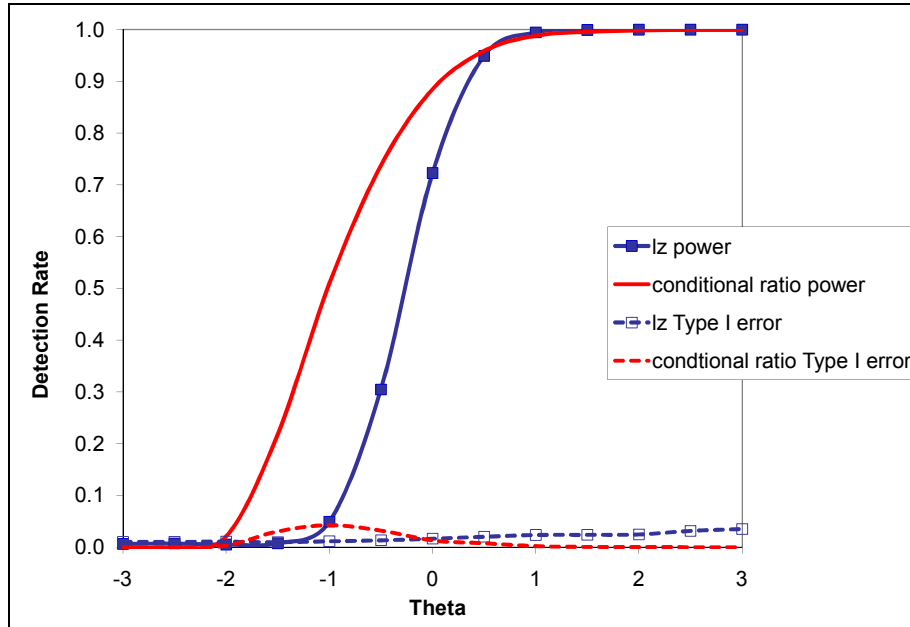


Figure 2: *Detection rates conditional on true θ for examinees who responded randomly to the previous 10 items and Type I errors for examinees who applied solution behavior to the 10 previous items.*

Estimated Theta. With empirical data, θ is unknown and must be estimated. The first 20 items were used to estimate an EAP θ , using a standard normal prior. The correlation with the true score in the null distribution, with cases weighted by the standard normal density, was .91 for a reliability of .83. The mean estimate was -0.003 and, as would be expected for EAP scores, the estimates were biased towards the mean and the standard deviation was 0.91.

In the null distribution, when some of the first 20 items were used in the calculation of ℓ_z the mean was greater than 0 (0.18 when all 10 of the items used in ℓ_z were among the first 20 used to estimate θ) and the standard deviation was less than 1 (0.88). This mean and standard deviation would be expected to make the test conservative, but there was still a negative skew in the distribution and the combined effect was a Type I error rate of .007, very close to the nominal

rate of .005. Beginning with item 30, when none of the items used for calculating ℓ_z were used for θ estimation, the mean was less than 0 (-0.11) and the standard deviation was greater than 1 (1.13). The Type I error rate for ℓ_z was .033 for these items, twice the error rate found with true θ . The corresponding conditional likelihood ratio cut-off was 4.538. This cut-off was based only on items after item 30, but for comparison purposes the same cut-off was applied to the earlier items. Also, the marginal likelihood ratio, with the likelihood of the response pattern averaged over the θ distribution as in Equation 2, was added for comparison. Because θ is not estimated for this index, it might be more accurate than using a fallible estimate of θ , especially when the θ estimate would otherwise be contaminated by some random responses. The cut-off value for the marginal likelihood ratio, also with a Type I error rate of .033, was 8.112.

Figure 3 provides the same information as Figure 1, but based on the estimated θ instead of the true θ . The results are reported separately depending on where students began responding randomly, because the first 20 items were used in θ estimation. When random responses began at item 11, the θ estimate was contaminated with random responses. When random responses began at item 21 (not shown), the θ estimate was not contaminated, but some of the items used to estimate θ were also used to calculate the fit indices. When random responses began at item 31 or later, none of the items used to estimate θ were used to calculate fit.

When random responding began after item 30, the power for ℓ_z and the conditional likelihood ratio were nearly identical to their power using true θ . Estimating θ , with a reliability of .83, had little impact on power. Not shown in the graph, when random responding began at item 21, power was lower initially (when more of the same responses were used both to estimate θ and to calculate fit) but by 10 random responses (item 30) power was essentially the same as it

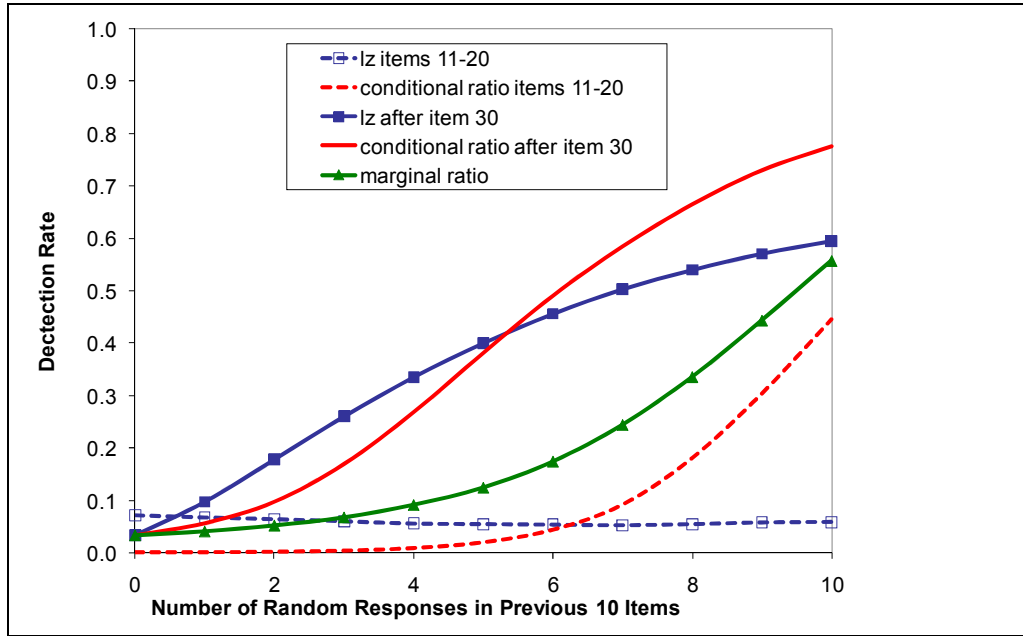


Figure 3: *Detection rates, based on estimated θ , as the number of random responses increases.*

was for 10 random responses later in the test. However, when random responding began at item 11 and contaminated the θ estimate, it was far less detectable. Power for ℓ_z was $< .1$; it was only slightly greater than the Type I error rate. Power for the conditional likelihood ratio increased with the number of random responses reaching .45 for 10 random responses – considerably lower than it was with true θ , but much higher than ℓ_z . Finally, the marginal likelihood ratio, which did not depend on where the random responses started and thus is shown as a single line in the graph, was more powerful at detecting random responses within the first 20 items but not as powerful as ℓ_z and the conditional ratio for random responses later in the test.

Figure 4 shows power for examinees who randomly responded to all 10 of the previous items, beginning sometime after item 30. The Type I error rate for the same sets of items is also displayed. When θ was known (Figure 2), examinees with low θ were virtually never flagged; both power and the Type I error rate were essentially zero because random responses were not

inconsistent with solution behavior. When θ was estimated, the estimates were too high for some of these low- θ examinees so both power and the Type I error rate increased. This effect was more noticeable for the conditional likelihood ratio than for ℓ_z . For moderate and high proficiency examinees, power was close to one. The marginal likelihood ratio had relatively constant power, as would be expected. Its power was higher than the other indices for low proficiency examinees but much lower than the other indices for moderate and high proficiency examinees. Further, its Type I error rate was very high for low proficiency examinees; their responses were aberrant for the population as a whole.

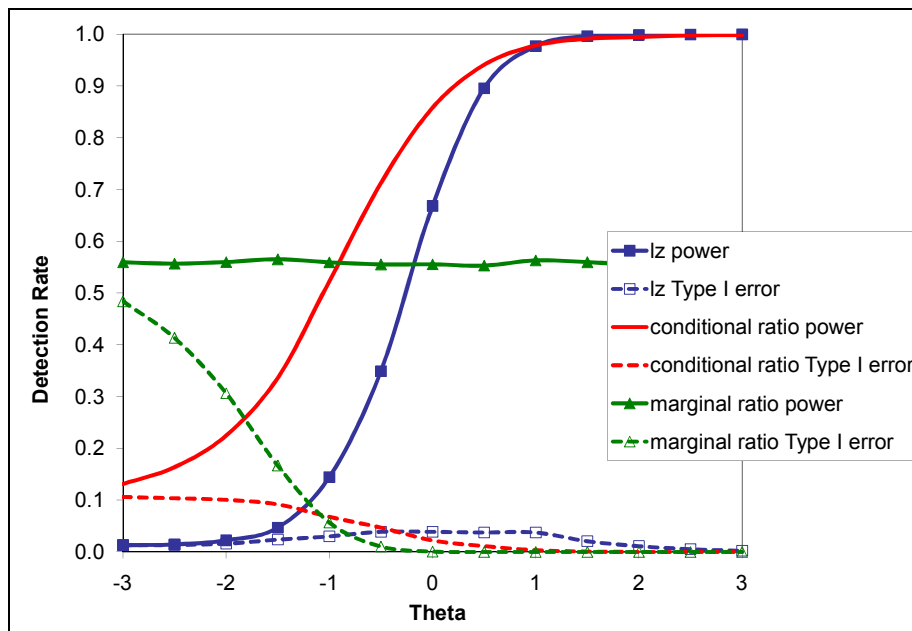


Figure 4: *Detection rates conditional on estimated θ for examinees who responded randomly to the previous 10 items and Type I errors for examinees who applied solution behavior to the 10 previous items.*

Examinee as the Unit of Analysis. These results have focused on the Type I error rate and power for each string of responses. If one's interest is focused on the examinee instead, the false-

hit rate is higher. When using a rolling fit index, each examinee has multiple opportunities to be flagged as misfitting. For example, with this 60-item test, if the fit was calculated for each item beginning at item 31 (the first point at which the fit index would not be based on any items used to estimate θ), each examinee would have 30 rolling fit values which correspond to 30 opportunities to be flagged as misfitting. Figure 5 shows the probability of being flagged at least once for examinees who began responding randomly at item 31 (power) and for examinees who maintained solution behavior throughout the test (false positives). Though the false positive rates show the same general pattern as in Figure 4, they are considerably higher at most θ values. However, most of the non-random (null distribution) examinees who were flagged were only flagged a few times. Figure 6 shows the average number of times, out of 30, that each group of examinees was flagged. Examinees with middle or higher θ s were flagged most of the time; the first few items to which they responded randomly were often not flagged because they responded to previous items in the set of 10 with solution behavior. Examinees who responded with solution behavior were flagged only about 1 time on average. These findings suggest that, if the focus is on the examinee, more complex flagging rules based on the total number of times flagged would help reduce the overall examinee-level false positive rate.

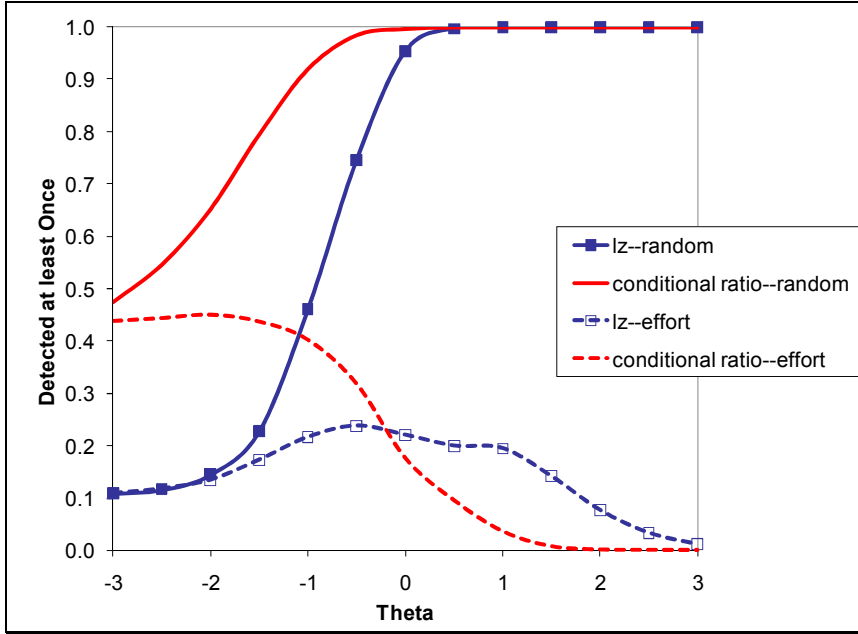


Figure 5: Total detection rates conditional on estimated θ for examinees who responded randomly to items 30-60 and total false positive rates for examinees who applied solution behavior to all items.

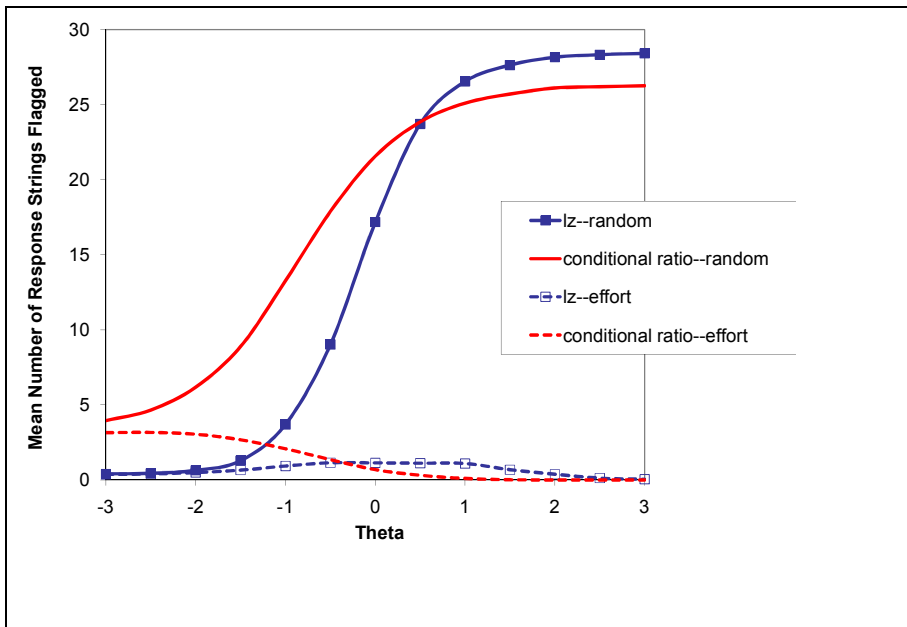


Figure 6: Mean number of misfitting values for examinees who responded randomly to items 20-60 and for examinees who applied solution behavior to all items.

Study 1 Discussion

The conditional likelihood ratio was more powerful than ℓ_2 for detecting misfit due to random responses to 6 or more items in a string of 10 items. The poor power for detecting random responding among low-proficiency examinees does not present a practical problem. The responses of these examinees are no more likely to be correct when they use solution behavior, so their score estimates are not greatly affected by random responses. They would have little impact on the group means, reliability, and validity.

Estimating θ from the first 20 items was effective for detecting random responding later in the test, though power decreased considerably if random responding began in the first 20 items. The marginal likelihood ratio would be more powerful in this context because it does not depend on the proficiency estimate, but at the cost of flagging too many examinees who were using solution behavior but were producing random responses due to low proficiency. In many contexts, most examinees would begin with a reasonable amount of effort and switch to random responding only later in the test, so the conditional likelihood ratio would have the highest detection rate overall and especially among high-proficiency examinees, where random responding would have the largest impact on scores.

When the examinee was the unit of analysis, the false positive rate was higher because there were many occasions for misfit to be detected. One approach might be to develop more complex rules, such as flagging an examinee only if at least 5 of the examinee's rolling fit values were flagged as misfitting. However, this would lose the advantage of calculating fit in real time and targeting the location(s) in the test where the examinee responded randomly. If the examinee is the unit of interest, it would be simpler to target longer strings of responses for the rolling fit, perhaps the last quarter or third of the test where examinees are more likely to give up. Again,

such an approach would not allow for the detection of short strings of random responses within the middle of the test.

Study 2: Real Data

Data from the ETS Major Field Achievement Test (MFAT) in Business were used to explore if there was a relationship between RTE and fit. The dataset contained 120 items and 10,004 examinees. Due to the length of the test and the typically low-stakes context under which many universities administer it, some amount of non-effortful responding was expected (for more information on rapid guessing in this data, see Setzer, Wise, and van den Heuvel, 2008). The focus of the research was whether students with high levels of rapid-guessing (low RTE) during the test would be detected as random responders using the conditional likelihood ratio index.

Method

Item parameters were estimated through MML estimation using BILOG 3. EAP θ s were estimated from the first 20 items, and the conditional likelihood ratio, using the estimated θ and responses to the previous 10 items, was calculated for items 20-120. The conditional likelihood ratio was used as the index of fit for this part of the study because it was more sensitive to random responding than ℓ_z . As in the simulation, a rolling fit index was based on the previous ten items. For comparison, a rolling RTE was also based on the previous 10 items. Again, RTE is the proportion of items to which the examinee responded with solution behavior. The thresholds for solution behavior were selected by visual inspection of the response time distribution. For most items, there was a small spike in number of respondents in the first few seconds; the threshold was defined as the end of this first small spike. Most items had thresholds around 4 seconds, with a range of 2-10 seconds.

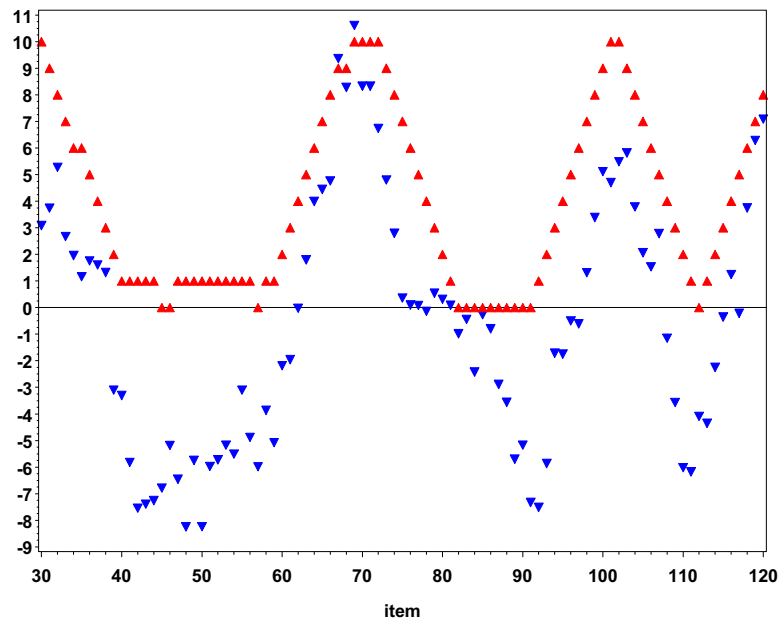
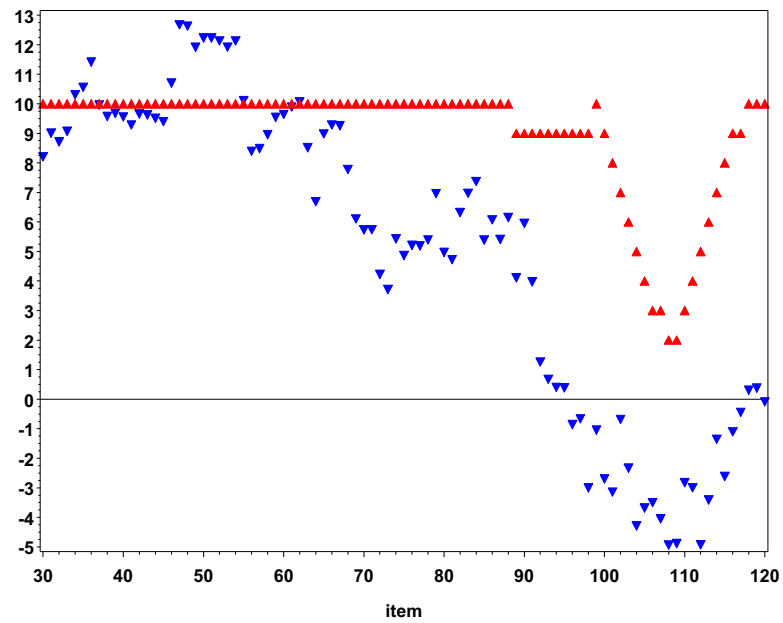
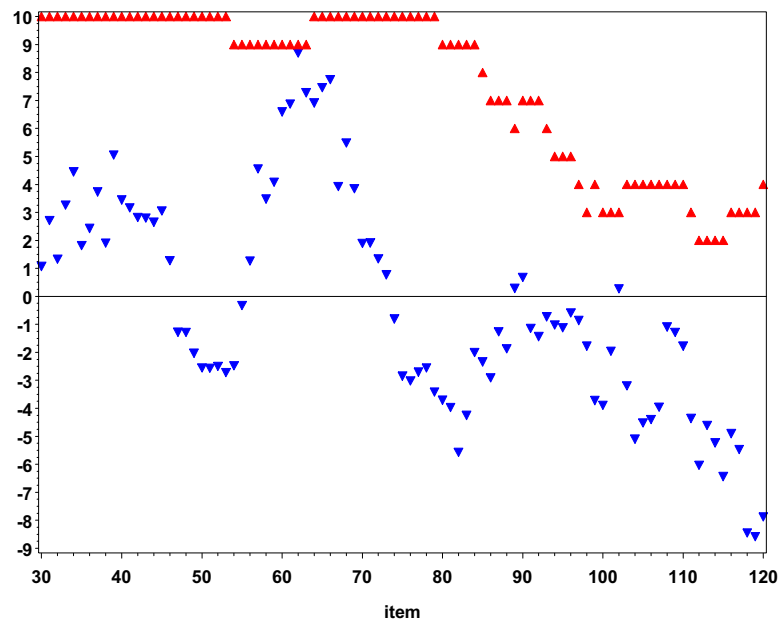
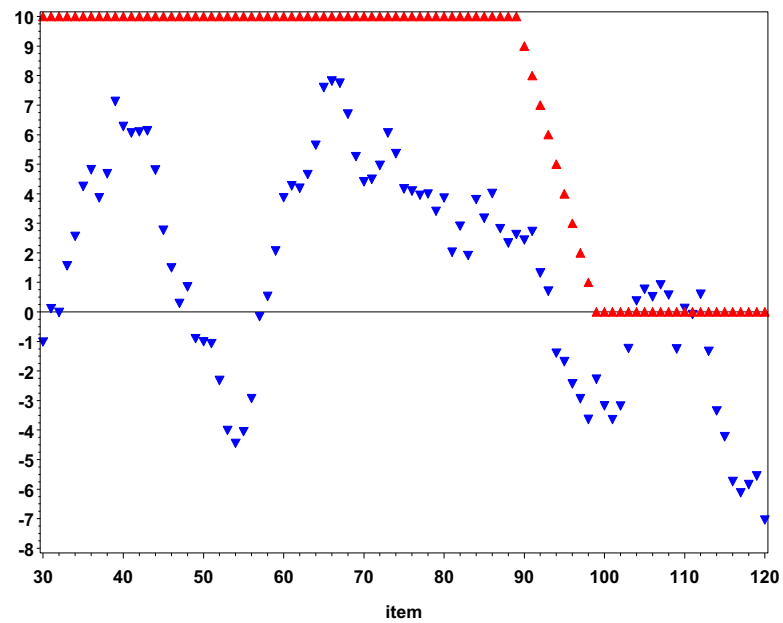
For comparison, item responses were simulated for a null (fitting) distribution, using the item parameter and θ estimates from the ETS data as the generating parameters (θ estimates were first unbiased by dividing by the estimated reliability).

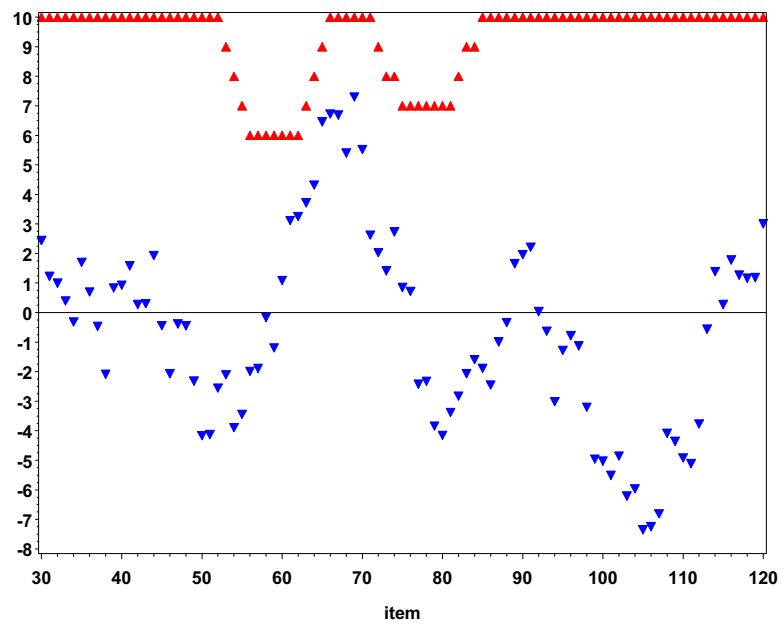
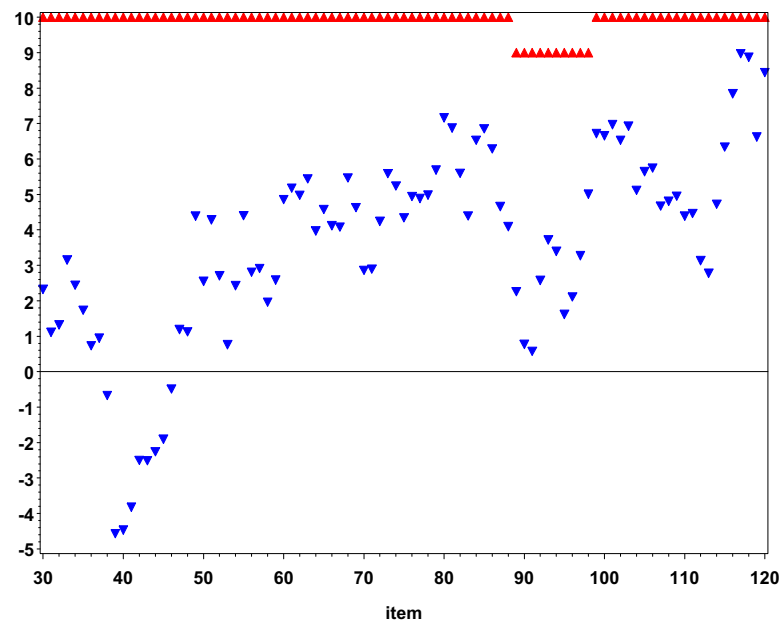
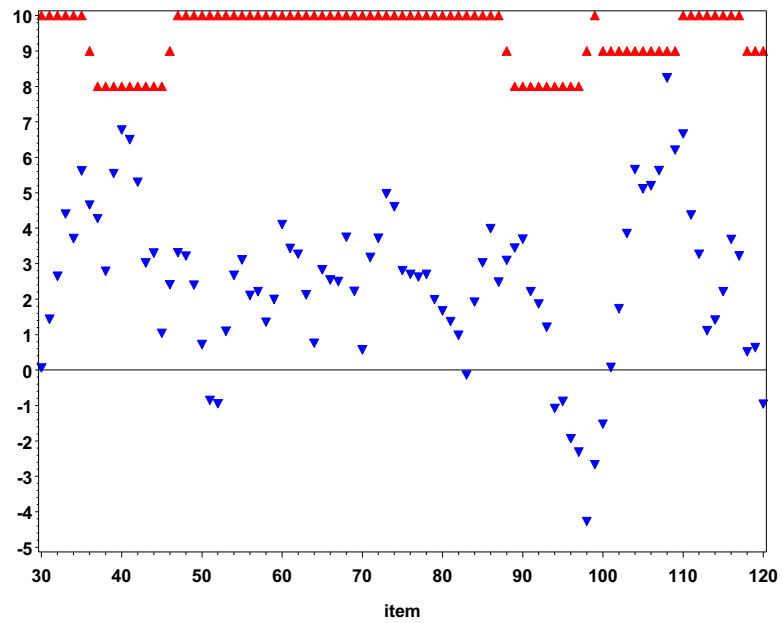
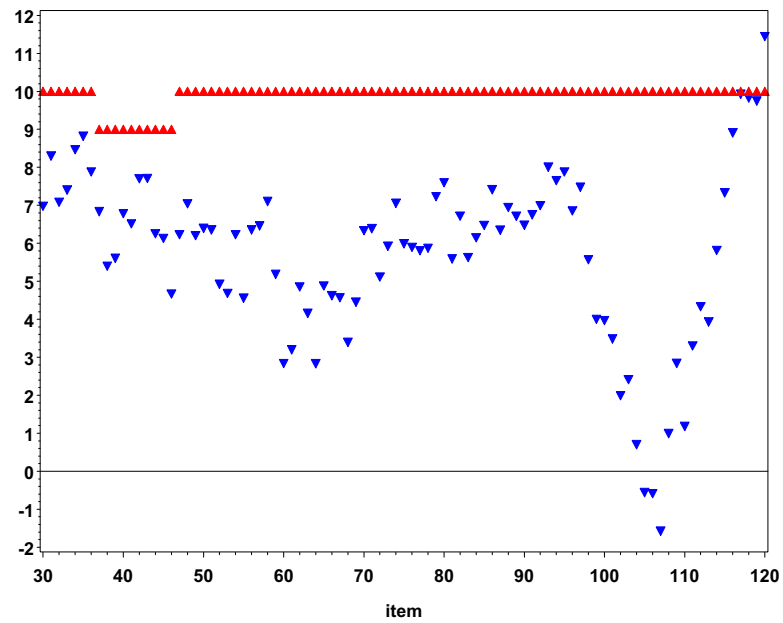
Results

To summarize the relationship between RTE and fit, a within-person correlation was calculated between the negative natural log of fit and RTE. The natural log was used to reduce the skew in the ratios (the natural log was close to normally distributed), and the negative was used so that patterns which fit the random response model better than the solution behavior model would have lower values. The correlation could only be calculated for the 2086 examinees who made at least one rapid-guess (as determined by the item threshold). On average, this within-person correlation was only .06. In Study 1, power was low for low values of θ , so a high correlation would not be expected in that range. Therefore, the correlation was also calculated only for $\theta > 0$. Within this range, the correlation increased slightly, to .14. Clearly, RTE and fit provide different information for most examinees².

The relationship between RTE and the negative log of the conditional likelihood fit ratio was plotted within each examinee to search for patterns. Figures 7 and 8 show the fit and RTE for a few selected examinees. In Figure 7, there seems to be a strong relationship between fit and

² Residual response time was also tested as a possible correlate with fit. The residual response time was the proportion by which the examinee's response time on the 10 items up to item i was slower than the mean response time on these items, minus the proportion by which the examinee's response time on the first 20 items was slower than the mean response time on the first 20 items. Thus an examinee who was 10% slower than other examinees on both the current item subset and the items at the beginning of the exam would have a residual of zero. The average within-person correlation of this index with log-fit was .02 for the population and .06 for the restricted sample with $\theta > 0$.





RTE; for higher values of RTE, the fit is better. But in Figure 8, there seems to be no relationship or a negative relationship. Visual examination of these within-person plots suggested that examinees with low or negative correlations tended to have generally high and stable RTEs, with only one or a few items meeting the rapid guess threshold. A single rapid guess decreases the rolling RTE to .9 for 10 items and often these periods did not correspond to poor fit. In short, examinees who had periods of very low RTE tended to have periods of poor fit at the same point as the very low RTE (if θ estimated from the first 20 items was at least -0.5). But many examinees who did not have very low RTE also had some periods of poor fit.

To explore this issue of very low RTE, 128 examinees were identified whose rolling RTE sunk to .1 or 0 at some point during the test. Essentially, these examinees abandoned test effort. Of these, 78 had estimated θ s (based on the first 20 items) < -0.5 ; as shown in Study 1, the random responses would not be misfitting for these examinees³ so they were not examined further. Among the remaining 50, all had at least one very poor rolling fit value; 48 of the 50 had at least one log fit ratio lower than -3 (the random response model was about 20 times as likely as the solution behavior model). However, so did about 25% of the examinees whose rolling RTE never dropped below .80 (and $\theta > -0.5$ for comparability). It is possible that this simply indicates that many random responders do not respond rapidly enough to meet the rapid-guess threshold. To check this explanation, data were simulated to follow the solution behavior model using the estimated item parameters for this ETS data set. In this simulated data set, 29% of the

³ There are two reasonable explanations for why so many low-RTE examinees had low θ estimates. Students who genuinely have low levels of knowledge may become discouraged and resort to rapid-responding to get through the test. Or the θ estimates may be artificially low because these students did not try very hard even on the first 20 items on the test.

examinees had at least one rolling fit value at this level, so the rate seen among examinees with high RTEs seems to be due to chance, not to random responding that is not rapid.

To target fewer high-RTE examinees, various rules were tried for targeting examinees only with several consecutive poor fit values, but these rules were not as successful at detecting all of the examinees with very low RTE. Finally, the fit and RTE plot within each of these examinees was visually inspected. For each of these examinees, the periods of poor fit corresponded to the periods of low RTE (see Figure 7 for examples).

Study 2 Discussion

Except for the examinees with low θ values, all examinees with a very low rolling RTE also had quite poor rolling fit (random model fit much better than solution-behavior model). Moreover, the low rolling RTE and poor rolling fit coincided at the same items. However, examinees with high RTE sometimes had poor rolling fit values as well. In other words, low RTE led to poor fit, but poor fit did not always indicate low RTE. One possible explanation is that the measures tap low effort in slightly different ways. However, examination of the fit in the simulated data showed poor rolling fit was found by chance alone about as often as it was found among the high RTE examinees. Thus, poor fit among examinees with high RTEs generally seems to be a Type I error rather than a good indication that the examinee is responding randomly but not rapidly. At any one point in the test, probability of a Type I error was low (defined as 2%), but given the length of the test there was a larger probability of at least one Type I error. Given the low base rate of low RTE, an examinee flagged for misfit was more likely to a false positive than a true positive (with a true hit defined as an examinee with very low RTE). The rolling fit index correctly identified nearly all of those with low RTE, but most of those identified did *not* have

low RTE. Rolling fit had a high sensitivity to low RTE, but the specificity was lower than desired.

Summary and Conclusions

The rolling conditional likelihood ratio and the rolling ℓ_z each have reasonable power for detecting examinees who begin responding randomly to a string of items, unless the examinees exhibit low proficiency in their responses to the items at the beginning of the test. Power is greater for the conditional likelihood ratio than for ℓ_z . Failure to detect random responses among low-proficiency examinees is reasonable because random responses would not be aberrant for these examinees. It is only problematic if the exhibited low proficiency is due to lack of effort even at the beginning of the test rather than to low levels of knowledge. But if one is interested in distinguishing between low scores due to low proficiency and low scores due to lack of effort, using the likelihood of the exact response options, rather than just the pattern of rights and wrongs might be helpful. This approach has been used with ℓ_z (Drasgow & Levine, 1985; Drasgow, Levine, & Williams, 1985). In the context of detecting random responding, this could be helpful if some response options are chosen far less often than random by low-proficiency examinees attempting solution behavior.

If a test administrator were interested in flagging misfitting examinees rather than misfitting responses, the overall Type I error rate for an examinee would increase because the index is calculated many times for each examinee. To reduce this error rate, longer fit intervals or more complex flagging rules, such as total number of times flagged or number of consecutive poor fit values, could be used. However, this would diminish the purpose of using a rolling index rather than a single index based on the items at the end of the test: the rolling index can identify

approximately where the random responses begin and can identify examinees who respond randomly for a period and then begin responding with solution behavior again.

In the real data, almost all students with low rolling RTE and at least moderate proficiency had poor rolling likelihood ratio values. Because of the low base rate of low RTE, most of the students with poor fit had no periods of low RTE. Again, rolling fit can not be recommended for use at the examinee level. However, the fact that periods of very low RTE corresponded with poor fit helps to validate the interpretation of low RTE. Based on the fit values, all of the students with very low RTE and at least moderate proficiency switched to a random response strategy. This information refutes alternative explanations for low RTE: these students did not start responding exceptionally rapidly because they had prior knowledge or were able to process information much faster than expected. If these explanations were accurate, the students' responses would not have fit the random response model so much better than the solution behavior model. In other data sets, these explanations may sometimes be true and rolling fit can be used to check whether low RTE corresponds to random responding.

In summary, rolling fit alone is not recommended for use at the examinee level because of the cumulative risk of false positives. However, rolling fit is useful for helping to understand more fully the rapid responses of an examinee. If the periods of rapid response correspond to periods of random response, as they did with the data examined in this study, then the responses are untrustworthy as indicators of the examinee's proficiency.

References

- DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*, 23-45.
- Drasgow, F. & Levine, M. V. (1985). Optimal detection of inappropriate test scores. Arlington, VA: Office of Naval Research, Personnel and Training Research Programs Office. (ERIC Document Reproduction Service No. ED265216)
- Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.
- Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9*, 47-64.
- Drasgow, F., Levine, M. V., & Williams, E. (1985) Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23(1)*, 17-27.
- Kong, X. J. (2007). *Using response time to investigate the effects of rapid guessing on the estimation of item and person parameters*. Unpublished doctoral dissertation, James Madison University.
- Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006, April). *Motivational effects of praise in response-time based feedback: A follow-up study of the effort-monitoring CBT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.
- Levine, M. V. & Drasgow, F. (1984). Performance envelopes and optimal appropriateness measurement. Washington, DC: Office of Naval Research, Psychological Sciences Division. (ERIC Document Reproduction Service No. ED263126)
- Levine, M. V. & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*, 161-176.
- Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.
- Molenaar, I. W., & Hoijsink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*, 75-106.
- Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121-129.

- Reise, M. L. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213-229.
- Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED383742)
- Schnipke, D. L. (1996, April). *How contaminated by guessing are item parameter estimates and what can be done about it?* Paper presented at the annual meeting of the National Council on Measurement in Education, New York. (ERIC Document Reproduction Service No. ED400276)
- Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In Mills, C. N., Potenza, M.T., Fremer, J. J., & Ward, W. C. (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Mahwah, NJ: Lawrence Erlbaum Associates.
- Setzer, J. C., Wise, S. L., & van den Heuvel, J. R. (2008, March). An investigation of examinee test-taking effort on the Major Field Test in Business. Paper presented at the annual meeting of the American Educational Research Association, New York
- Wise, S. L., Bholra, D., & Yang, S. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice 25*(2), 21-30.
- Wise, S. L., & DeMars, C. E. (2005). Examinee motivation in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1-18.
- Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*, 19-38.
- Wise, S. L., & DeMars, C. E. (in press). A clarification of the effects of rapid guessing on coefficient alpha: A note on Attali's Reliability of Speeded Number-Right Multiple-Choice Tests. *Applied Psychological Measurement*.
- Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*, 163-183.
- Wise, S. L., & Smith, L. F. (in press). A model of examinee test-taking effort. In J. Bovaird (Ed.) *Contemporary issues in high stakes testing*. Washington, DC: American Psychological Association.
- van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327-345.