Modeling Answer Changes on Test Items

Wim J. van der Linden
CTB/McGraw-Hill

Minjeong Jeon
University of California, Berkeley

The probability of test takers changing answers upon review of their initial choices is modeled. The primary purpose of the model is to check erasures on answer sheets recorded by an optical scanner for numbers and patterns that may be indicative of irregular behavior, such as teachers or school administrators changing answer sheets after their students have finished the test or test takers communicating with each other about their initial responses. A statistical test based on the number of erasures is derived from the model. Besides, it is shown how to analyze the residuals under the model to check for suspicious patterns of erasures. The use of the two procedures is illustrated for an empirical data set from a large-scale assessment. The robustness of the model with respect to less than optimal opportunities for regular test takers to review their responses is investigated.

Keywords: answer changes; cheating detection; erasure analysis; response modeling

Introduction

The topic of test takers changing their initial answers on a test has intrigued test theorists for quite some time. Originally, the interest was mainly motivated by the question of whether or not test takers should be admitted—or even encouraged—to review their initial responses on test items and change them when they have second thoughts. The answer to this question, of course, depends on whether test takers can be expected to profit from such reviews. More specifically, we should know how the expected change of scores relates to the ability of the test takers and the properties of the items.

The general conclusion from a long tradition of empirical research on this question has been that, contrary to the popular belief that first impressions are always best and test takers should stick to their initial answers, positive but minute gains can be expected for test takers with moderate to high abilities but that there are no positive gains for test takers with low abilities (e.g., Lynch &
Smith, 1972, April; McMorris, DeMers, & Schwarz, 1987; Schwarz, McMorris, & DeMers, 1991). However, each of these earlier studies was mainly interested in the marginal probability of the benefit of answer changes for a population of test takers. A recent study based on item-response modeling that analyzed the benefits conditionally on the test taker’s ability level confirmed the popular belief and showed rather dramatic loss at each ability level (van der Linden, Jeon, & Ferrara, 2011).

The interest in the question has been rekindled by the introduction of computerized adaptive testing. Changing earlier responses in an adaptive test leads to the necessity of reestimating the current ability estimate and, hence, to loss of efficiency because of less than optimal selection of the items that have already been administered after the item for which the response was changed. For adaptive testing, the critical question, therefore, is whether the expected benefits of score changes are enough to offset the associated decrease in score accuracy. Studies addressing this issue are found, for instance, in Bowles and Pommerich (2001, April), Ferrara et al. (1996, April), and Wise (1996, April).

The current interest in answer changing was motivated by the problem of cheating on educational tests. Several types of cheating involve a change of the initial answers on the answer sheet. The prime example in this research is teachers or school administrators erasing incorrect answers on answer sheets and replacing them with correct answers after their students have finished the test. But the proposed erasure analyses may also help detecting test takers communicating their answers with each other and making changes once they have worked through the items for the first time.

The possibility of schools cheating on state assessments has received considerable attention through the work of Jacob and Levitt (2003a, 2003b, 2004) and subsequent popularization in Levitt and Rubner (2005). Their method, which is described in Jacob and Levitt (2003a, appendix), consisted of the detection of unlikely patterns of response alternatives on multiple-choice (MC) items using multinomial logits regressed on the students’ scores on assessments in the previous and next years as well as information on their background. Thus, the method essentially evaluated the choice of the response alternatives against the probabilities for similar students. Blocks of unexpected correct alternatives on answer sheets for students with the same teacher were then inspected more closely for possible cheating.

Their method can be improved in the following two ways, though. First, the method misses important information because it is based on the final responses rather than directly on the observed erasures. Modern optical scanners can easily be set to automatically detect answers that have been erased and replaced by new answers. The information tells us if an erasure was from a right to wrong (RW), wrong to another wrong (WW), or wrong to the right (WR) answer. Of course, if the focus is on the detection of cheating, the interest will be primarily in the potentially suspicious WR erasures. The analyses of the numbers and patterns...
of such erasures may reveal important information about irregular response behavior when this has occurred.

Second, regular test takers may occasionally produce unusual numbers or patterns of erasures as well. The only possible way to distinguish statistically between regular and irregular behavior is to have a satisfactory model for the probabilities of regular erasures on the items and evaluate the answer sheets under these probabilities. However, such a model can only hold when it adequately accounts for the rather complicated interaction between the test taker’s ability measured by the test and the probability of an erasure. For instance, a more able test taker is less likely to produce an incorrect initial answer. However, given an incorrect answer, he or she is more likely to make a WR rather than a WW erasure. The model should thus distinguish between the probabilities of initial answers and subsequent erasures, and condition them on the test taker’s ability.

This paper proposes a probability model that does so and introduces a statistical test derived from it with the number of WR erasures as the test statistic. Besides, it is shown how the model can be used for an analysis of the patterns of residual WR erasures. The use of the detection procedures will be illustrated for an empirical data set from a recent large-scale assessment. We also present results from a study of the robustness of the model with respect to an assumption about the testing time on which it rests, and discuss how the procedure should be implemented when the assumption is violated.

The proposed test should never be used blindly or relied on as the only source of evidence of cheating. Although it has the full rigor of a statistical test, it does make Type I and II errors—sometimes with an obvious explanation; for example, a test taker who goes through the whole test quickly with the intention to capitalize on the later review to correct earlier mistakes due to sloppiness. Several ways of dealing with such errors are available, though. For instance, as a rule, once a set of answer sheets is flagged, we should follow up with more qualitative analyses of their patterns of erasures. The method of residual erasures presented below can be used to search for common blocks of unlikely WR erasures for the same teachers, unlikely erasures at the end of the test, pairs of test takers with matching erasures that sat next to each other, and so on. Also, a statistical test of cheating should always be used as part of a larger suite of procedures that use independent sources of information, for instance, the responses on the test items (Holland, 1996; Lewis & Thayer, 1998; Sotaridona, van der Linden, & Meijer, 2006; van der Linden & Sotaridona, 2004, 2006), the response times (van der Linden, 2009; van der Linden & Guo, 2008), external evidence (van der Ark, Emons, & Sijtsma, 2008), as well as the source addressed in this research—erasures found by the optical scanner. Finally, the current authors are much in favor of a strategy already adopted by several testing agencies to follow up on damaging evidence not by incrimination but with the announcement that the test could not be scored because of statistical irregularities and the candidate is offered the opportunity to retake the test (under strictly proctored conditions, of course).

182
Model Formulation

Before presenting the model, we will discuss its underlying conceptualization. The model is for a test with MC items administered in a paper-and-pencil format, with all items already carefully pretested and calibrated under the response model in use for the program; for instance, the three-parameter logistic (3PL) model or one of the mainstream polytomous response models.

Basic Conceptualization

A key assumption on which the development of the model rests is that the test takers have enough time to answer all items in the test and review their answers once they are done. The case of tests with too limited time is dealt with later in this paper.

The regular response process thus has two different stages:

1. A first stage consisting of the production of the initial responses. We use \( U_{ni}^{(1)} \) to indicate the initial responses by test takers \( n = 1, ..., N \) to items \( i = 1, ..., I \), where \( U_{ni}^{(1)} = 1 \) is a correct response and \( U_{ni}^{(1)} = 0 \) an incorrect response. As the items are operational and have been pretested to fit the response model in the testing program, the probability of observing response \( U_{ni}^{(1)} = u_{ni} \) can be safely assumed to be represented by this model.
2. A final stage in which the test takers review their responses and make whatever changes necessary. The results are the final responses \( U_{ni}^{(2)} \). The focus of this study is on the modeling of the probabilities of these responses.

Different types of score patterns for the two stages are possible. With respect to the occurrence of an erasure, three different types of patterns should be distinguished:

1. \( (U_{ni}^{(1)} = 1, U_{ni}^{(2)} = 0) \), which occurs when an initial correct answer is erased and replaced by an incorrect answer. We will refer to this result as an RW erasure.
2. \( (U_{ni}^{(1)} = 0, U_{ni}^{(2)} = 0) \), which occurs when an initial incorrect option is replaced by another incorrect option. The result is a WW erasure.
3. \( (U_{ni}^{(1)} = 0, U_{ni}^{(2)} = 1) \), which occurs when an incorrect initial answer is replaced by a correct answer. The result is the potentially suspicious WR erasure.

Notice that erasures with one correct option replaced by another (RR erasures) are excluded for the typical case of MC items with one correct alternative.

However, we will observe cases where the initial response is correct and no erasure on the answer sheet has been made. As the test taker is assumed to review all responses, we should interpret this case as a review that has led to a confirmation of the initial response. It will therefore be scored as \( (U_{ni}^{(1)} = 1, U_{ni}^{(2)} = 1) \).
When we want to emphasize the fact that a direct observation is missing, this case will be denoted as \( (U_{ni}^{(1)} = 1, U_{ni}^{(2)} = *) \).

Likewise, the case of an initial incorrect response that is not changed can be interpreted as a test taker reviewing the response and confirming it; that is, as \( (U_{ni}^{(1)} = 0, U_{ni}^{(2)} = 0) \). If we want to emphasize an actually missing observation, we will denote this case as \( (U_{ni}^{(1)} = 0, U_{ni}^{(2)} = *) \).

The assumption of only one review of an initial response is somewhat arbitrary. In principle, test takers may go through more than one revision of some of their responses and even erase more than one response for an item. The probability of this happening will be considered as negligible, though. For convenience, we will therefore use the recorded erasures as the only erasures, regardless of the number of review steps used to reach it.

**Models**

For the initial stage, we assume the regular response model that holds for the testing program. We will first discuss the case of the 3PL model and then deal with polytomous alternatives. The response function for the 3PL model is

\[
\Pr\{U_{ni}^{(1)} = 1\} = c_i + (1 - c_i) \frac{\exp[a_i(\theta_n^{(1)} - b_i)]}{1 + \exp[a_i(\theta_n^{(1)} - b_i)]},
\]

where \( \theta_n^{(1)} \) is the ability of test taker \( n \) and \( a_i, b_i, \) and \( c_i \) are the discrimination, difficulty, and guessing parameters of item \( i \). As already indicated, the item parameters are assumed to be estimated with satisfactory precision and the fit of the model to be evaluated prior to the operational use of the items.

Obviously, the result of the review of an item depends on the initial response. Hence, the usual assumption of conditional independence between the initial and final responses (“local independence”) does not hold. We therefore model the probability distributions of the final responses conditional on the initial responses; that is, consider different response probabilities for \( U_{ni}^{(2)} = 1 \) given \( U_{ni}^{(1)} = 0 \) and \( U_{ni}^{(1)} = 1 \).

The basic idea is to model each of these conditional probabilities using a version of the initial 3PL model with the following choice of parameters:

1. Each of the two conditional models has free item parameters \( a_i \) and \( b_i \) to allow for the different impact of a correct and incorrect response in the initial stage on the change of the response probabilities.
2. It does not make much sense to assume test takers guessing on any of the items during their review of their initial responses. If they have guessed initially, there is no need for them to guess on the same item again. Likewise, if the initial response was not obtained by guessing, it seems unreasonable to assume that they...
improve on it by guessing as yet. We therefore set the final-stage parameters $c_i = 0$. Thus, the conditional models will be 2PL models.

(3) As the free $a_i$ and $b_i$ item parameters already absorb all the differences between the two stages at the level of the individual items, it is not necessary to have free $\theta_i$ parameters for the final stage. We therefore assume that the final-stage ability parameters are equal to their initial values; that is, $\theta_n^{(2)} = \theta_n^{(1)}$.

This last assumption is based on the fact that, for the 3PL model, only the difference between the $\theta_n$ and $b_i$ parameters is identified. A common change in the $\theta_n$ parameters is therefore formally equivalent to the opposite change in the $b_i$ parameters. However, leaving the $a_i$ and $b_i$ parameters free and fixing the $\theta_n$ parameters, rather than the other way around, allow for possible additional changes at the level of the individual items as well. Also, note that, because of the same lack of identifiability, when both the item and the examinee parameters would have been left free, a comparison between their estimates in the models for the two stages as in Figure 1 below would have been impossible. Finally, fixing the $\theta$s rather than the item parameters gives important advantages as to the choice of the statistical software we can use for the estimation of the model parameters (see below).

The final-stage models can thus be written as

$$
\Pr\{U_n^{(2)} = 1|U_n^{(1)} = 1\} = \frac{\exp[a_1i(\theta_n^{(1)} - b_{1i})]}{1 + \exp[a_1i(\theta_n^{(1)} - b_{1i})]},
$$

and

$$
\Pr\{U_n^{(2)} = 1|U_n^{(1)} = 0\} = \frac{\exp[a_0i(\theta_n^{(1)} - b_{0i})]}{1 + \exp[a_0i(\theta_n^{(1)} - b_{0i})]}.\tag{3}
$$

Observe that the probability in Equation 3 is the probability of a WR erasure. As this is the suspicious type of erasure in the intended cheating analysis, the remainder of the paper will focus on the estimation of this probability. Likewise, the complement of Equation 2 is the probability of an RW erasure. However, the complement of Equation 3 is not the probability of a WW erasure but the compound event of a WW erasure or a test taker confirming an incorrect initial response.

Polytomous models. The previous results immediately suggest how to deal with testing programs for which a polytomous response model is in use: the (generalized) partial credit model, graded response model, or nominal response model, say. For each of these models, the final-stage models are conditional versions, given the initial response with free item parameters but the ability parameters fixed at their initial values. As these models have no guessing parameters, there is no need to set such parameters to any specific value.
Observe that for the polytomous case, we always have more possible outcomes of the review than the earlier three for the dichotomous case (RW, WW, and WR). For instance, for an item with three alternatives, there are two possible ways of changing an incorrect response into a correct response. The same holds for the reverse change. And there are two possible ways of replacing an incorrect response by another incorrect response. Finally, a test taker can confirm each of the three options. In total, we thus have to consider nine different probabilities.

Although cheating detection based on a polytomous response model would be more refined, the focus in this paper is on the analysis of suspicious changes in the correctness of the answers. We therefore refrain from a further elaboration of the polytomous case.

**Statistical Treatment**

The probabilities in Equations 2 and 3 are estimated from two separate subsets of the final responses, $U_{ni}^{(2)}$: one subset for which the initial response was $U_{ni}^{(1)} = 0$ and the other for $U_{ni}^{(1)} = 1$. These complementary subsets thus have structurally missing data. The exact amount of missing data depends on the location of the ability distribution of the test takers relative to the difficulties of the items. If the difficulty of the test matches the ability distribution well, the percentage of missing data for the estimation of the probabilities of the WR erasures in Equation 3 can be expected to be close to 50%.

Although missing data means lack of information, it helps that the guessing parameter can be set equal to $c_i = 0$ for all items and we already have estimates of the ability parameters, $\theta_{ni}^{(1)}$. Instead of fitting a 3PL model with unknown item and person parameters, we only have to estimate the $a_{0i}$ and $b_{0i}$ parameters. For one thing, this means we avoid the poor identifiability of the $a_{0i}$ parameters due to the presence of $c_i$ as a free parameter, which sometimes hampers the estimation of the 3PL model for small data sets. More importantly, the estimation problem is now equivalent to the estimation of the slope and intercept parameters in a simultaneous logistic regression analysis of the responses $U_{ni}^{(2)}\mid U_{ni}^{(1)} = 0$ on the ability estimates $\theta_{ni}^{(1)}$ for each of the items. Treatment of the estimation problem in this regression framework requires the reparameterization

$$\Pr\{U_{ni}^{(2)}\mid U_{ni}^{(1)} = 0\} = \logit^{-1}(b_{0i}^* + a_{0i}^* \theta_{ni}^{(1)}).$$

(4)

as well as back transformation of the estimated intercept parameters to $b_{0i} = -a_{0i} b_{0i}^*$. We adopt the framework in the remainder of this section.

As the data for $U_{ni}^{(2)}$ given $U_{ni}^{(1)} = 0$ are potentially sparse, cases of logistic regression with quasi-complete separation (except for a few overlapping responses, each of the two possible responses corresponds with unique levels of the predictor variable) or even complete separation (no overlap at all) may
occur. To avoid unstable or even unbounded slope estimates due to such conditions, a Bayesian approach with weakly informative priors on the parameters helps. In the empirical studies later in this paper, we followed the suggestion for the priors for logistic regression introduced in Gelman, Jakulin, Grazia, Pittau, and Su (2008) to deal with possible separation. The priors for the slope parameters $a_{0i}'$ and intercept parameters $b_{0i}'$ are independent Cauchy distributions ($t$-distributions with 1 df) with location zero. The prior distribution for the slope parameters has scale 2.5; the one for the intercepts has scale 10. (These priors are to be imposed after a standardization of the scale of the predictors.)

The two priors are very weak indeed. The one for the slope parameters can be shown to be slightly less informative than the observation of one-half of a success and one-half of a failure in a Bernoulli trial with Equation 4 as probability for $a_{0i} = 0$ and $b_{0i} = 1$ (Gelman et al., 2008); the one for the intercept is even less informative. Nevertheless, as demonstrated extensively by these authors, these prior distributions offer just enough shrinkage to yield stable estimates for parameters with extremely small numbers of data while leaving the other parameter estimates essentially untouched.

Gelman et al. (2008) offer a convenient estimation procedure for these prior distributions in the form of an EM algorithm based on classical weighted least squares analysis from multilevel regression modeling, which quickly approximates the posterior modes and standard deviations. The procedure, which is available as the function bayesglm in the arm package for multilevel regression analysis in R, was used in all our examples.

The package is based on the standard logistic formulation,

$$\Pr\{y = 1\} = \logit^{-1}(X\beta).$$

Using indicator variables $d_i$ for the items $i = 1, \ldots, I$, the model in Equation 4 for test taker $n$ can be written as

$$\Pr\{U^{(2)}_{ni} | U^{(1)}_{ni} = 0\} = \logit^{-1}(b_{0i}'d_i + a_{0i}'d_i\theta^{(1)}_n),$$

with $\theta^{(1)}_n$ a known value. The representation in Equation 5 for test taker $n$ is obtained by the choice of the $n \times 2I$ design matrix

$$X_n = \begin{pmatrix} 1 & 0 & \cdots & 0 & \theta_n & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \theta_n & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & \theta_n \end{pmatrix},$$

and $2I \times 1$ parameter vector

$$\beta = (b_{01}', b_{02}', \ldots, b_{0I}', a_{01}', a_{02}', \ldots, a_{0I}').$$

For less sparse data, when the interest is not only in point estimates for the slope and intercept parameters but in the shape of their full posterior distributions, an
alternative in the form of MCMC sampling from the posterior distributions could be used; for instance, the Metropolis-Hastings within Gibbs procedure based on the probit approximation of Equation 3 in Johnson and Albert (1999, chap. 3). As we fix the final ability parameters to the estimates of $\theta_n^{(1)}$, an efficient approach would be to save the draws from the posterior distributions of $\theta_n^{(1)}$ from the initial stage and sample the vector with the saved draws while reestimating the item parameters during the second stage. An advantage of this alternative procedure is that it allows for the uncertainty in the initial $\theta_n^{(1)}$ estimates during the reestimation of the item parameters.

**Assumption of Full Review**

A key assumption of the model is that the test takers have enough time to answer the items in the test and review each of their answers. Ultimately, whether or not the assumption is fulfilled is a matter of appropriate planning of the test and instruction to the test takers. If the test does not give test takers enough time to review their answers to all of the items, they may still be able to review some of them.

We are able to directly check the assumption of full review in computerized testing, where the times spent initially and during revisits to the items are automatically logged. But for paper-and-pencil forms, we have to rely on appropriate assignment of time to tests that are not intended to be speeded.

If the test takers are not given enough time, violation of the assumption of full review leads to two potential types of confounded data:

1. It is no longer clear whether an unchanged correct response should be scored as $(U_{ni}^{(1)} = 1, U_{ni}^{(2)} = 1)$ or $(U_{ni}^{(1)} = 1, U_{ni}^{(2)} = *)$ for some of the items.
2. Likewise, it is unclear whether an unchanged incorrect response should be scored as $(U_{ni}^{(1)} = 0, U_{ni}^{(2)} = 0)$ or $(U_{ni}^{(1)} = 0, U_{ni}^{(2)} = *)$.

In case of less than full opportunity to review, always scoring unchanged correct responses as $(U_{ni}^{(1)} = 1, U_{ni}^{(2)} = 1)$ would generally lead to an overestimation of the probability in Equation 2 for the items that were not reviewed. This would not be a problem for the statistical test below because it is not based on these probabilities. Likewise, always scoring unchanged incorrect responses as $(U_{ni}^{(1)} = 0, U_{ni}^{(2)} = 0)$ would generally lead to an underestimation of Equation 3 for the items that were not reviewed. This would not be a problem either for the test takers who were unable to review these items because the test is based on their counts of WR erasures and no erasures whatsoever are recorded for them on these items. But it would be a problem for the occasional test taker who was still able to review these items and produced a legitimate WR erasure. In more statistical terms, a time limit that is too tight would give the detection procedures
below much more power in the sense that even a few erasures would be unlikely unless there was cheating. However, the price would be a Type I error for an occasional very able test taker who worked hard and was able to review the answers.

We will return to this issue in a robustness study with respect to violation of the assumption of full review later in this paper.

Detection of Cheating

We first present a statistical test of the null hypothesis of regular response behavior against the alternative of irregular behavior. As test statistic, we use the total number of WR erasures on all items for which the test taker has score \( U_{ni}^{(1)} = 0 \). For test taker \( n \), let \( i_n = 1, \ldots, I_n \) denote these items and \( E_n \) the total numbers of WR erasures.

The distribution of \( E_n \) is the result from a series of \( I_n \) independent Bernoulli trials, each with a different probability \( P_{ni} \equiv \Pr\{U_{ni}^{(2)} = 1|U_{ni}^{(1)} = 0\} \) given by Equation 3. The distribution is known as the generalized or compound binomial (e.g., Lord, 1980, sect. 4.1). It does not have a probability function in closed form but its probabilities can be calculated using the generating function

\[
\prod_{i_n=1}^{I_n} [Q_{ni} + zP_{ni}],
\]

where \( Q_{ni} \equiv 1 - P_{ni} \) and \( z \) is a dummy variable. Carrying out the multiplications in Equation 9 gives the probabilities of \( E = e \) as the coefficients of \( z^e \). As an example, the following table shows the probabilities for the case of \( I_n = 3 \):

Even for moderate values of \( I_n \), the number of products in these sums becomes prohibitively large, and the recursive procedure in Lord and Wingersky (1984) should be used. The procedure is simple and requires only a few lines of computer code. It begins with the case of one item and then iteratively adds a new item until the full test length is reached. For instance, for the case of \( I_n = 4 \) (i.e., the addition of a fourth item with probabilities \( P_4 \) and \( Q_4 \) to the above table for \( I_n = 3 \)), the probability of \( e = 2 \) is immediately obtained as the sum of \( P_1Q_2Q_3P_4 + Q_1P_2Q_3P_4 + Q_1Q_2P_3P_4 \) (row for \( e = 1 \) in the above table multiplied by \( P_4 \); that is, success on the extra item) and \( Q_1P_2P_3Q_4 + P_1Q_2P_3Q_4 + P_1P_2Q_3Q_4 \) (row for \( e = 2 \) in the above table multiplied by \( Q_4 \); that is, failure on the extra item).

The generalized binomial distribution is stochastically increasing in each of the probabilities \( P_{ni} \) (van der Linden & Sotaridona, 2006). Thus, for actual probabilities of WR erasures larger than the probabilities in Equation 3, the upper tail of the distribution of \( E_n \) is always to the right of the model distribution. Hence, for a test of level \( \alpha \), the critical value to flag the number of WR patterns by test
van der Linden and Jeon

Taker \( n \) as significant is the smallest value of \( E_n \) for which the area in the upper tail of the generalized binomial distribution is not larger than \( \alpha \); that is, the smallest value, \( e_n^* \), satisfying

<table>
<thead>
<tr>
<th>( e )</th>
<th>( \Pr{E = e} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>( Q_1Q_2Q_3 )</td>
</tr>
<tr>
<td>1</td>
<td>( P_1Q_2Q_3 + Q_1P_2Q_3 + Q_1Q_2P_3 )</td>
</tr>
<tr>
<td>2</td>
<td>( Q_1P_2P_3 + P_1Q_2P_3 + P_1P_2Q_3 )</td>
</tr>
<tr>
<td>3</td>
<td>( P_1P_2P_3 )</td>
</tr>
</tbody>
</table>

\[ \Pr\{E_n \geq e_n^*\} \leq \alpha. \quad (10) \]

Observe that, under the assumed model, except for discreteness and negligible item calibration error, this is an exact test in the sense of no large-sample approximation being involved. There is thus no need to check on the actual level of the test (e.g., through a simulation study). Besides, its power functions are identical to those for the test of a test taker copying answers from another test taker in van der Linden and Sotaridona (2006, figures 4–6). Although their statistics have different interpretations (number of erasures for one test taker versus number of matching responses between two test takers), both tests are based on the number of successes under the same family of generalized binomial distributions. The curves show that, for tests of only 20 to 30 items, the power to detect 5 to 10 irregular erasures would already be indistinguishably close to 100%.

As explained in the introduction, statistical tests of cheating should always be followed by more qualitative analyses to explain what may have happened. In the current context, this should be an analysis of the pattern of WR erasures for answer sheets that are flagged. We recommend an analysis in the form of a plot of the estimated WR residuals for the flagged test takers, that is,

\[ U^{(2)}_m - \Pr\{U^{(2)}_m = 1|U^{(1)}_m = 0\}, \quad (11) \]

as a function of the position of the items in the test. These plots are easy to interpret and immediately reveal suspicious patterns of WR erasures, such as unexpectedly large blocks of adjacent items with positive residuals, when they occur. An empirical example of the analysis is given in the next section.

**Empirical Example**

The procedure was applied to a data set from a large-scale assessment that consisted of the responses of 2,555 Grade 3 students to 65 mathematics items.
The items were part of a larger set that had been calibrated under the 3PL model using marginal maximum likelihood estimation with $\theta \sim N(0, 1)$ and shown excellent fit.

The answer sheets of the students were scanned using the option of erasure detection. The result was used to determine the $U_{ni}^{(1)}$ and $U_{ni}^{(2)}$ responses as indicated earlier. The $U_{ni}^{(1)}$ responses were used to estimate $\theta_n^{(1)}$ for each student in the data set using expected a posterior (EAP) estimation with a uniform prior over $[-4, 4]$.

The final-stage data set was split into subsets for $U_{ni}^{(1)} = 0$ and $U_{ni}^{(1)} = 1$, and the former was used to estimate the $a_{0i}$ and $b_{0i}$ parameters in the model for the WR erasures in Equation 3. Because of this conditioning on the initial responses, the set had some 60% of missing responses. This was no problem whatsoever for the bayesglm function with the priors described earlier, which converged quickly in all cases. As a measure of the overall fit of the model, we used pseudo $R^2$; that is, (null deviance-residual deviance)/residual deviance. For the current application $R^2 = .487$, which means that approximately 49% of the variance of the final responses was explained by the model.

Figure 1 compares the final estimates of the item parameters with the initial estimates. The final estimates of the discrimination parameters tended to be slightly higher than the initial estimates, but the estimates of the difficulty parameters were generally much higher. Both changes make sense. Because the guessing parameters $c_i$ were set equal to zero for the final stage, the slopes of the response functions for the items could be expected to be somewhat steeper. The

FIGURE 1. Comparison of the initial and final estimates of the discrimination ($a_i$) and item difficulty ($b_i$) parameters.
increase in the estimates of the difficulty parameters indicates that it is generally easier to produce a correct initial response than improving on an incorrect initial response during item review.

Next, the proposed cheating detection procedure was executed. For each student-item combination with the initial response $U_{ni}^{(1)} = 0$, the probability of a WR erasure was calculated from the estimates of the $a_{0i}$, $b_{0i}$, and $\theta_{ni}^{(1)}$ parameters. The probabilities were used to calculate the generalized binomial probabilities of the numbers of erasures, $E_n$, in Equation 9 for each student and to determine the critical values $e^*_{n, \alpha}$ in Equation 10 for $\alpha = .05$, .01, and .001. The percentages of students flagged as significant at these levels were 2.62, 1.29, and 0.47%, respectively. The minor differences between the percentages and the nominal levels of the tests can be explained by estimation error in the ability parameters. The size of the differences does not point at any massive irregularity during the administration of the test.

For the flagged students, we prepared plots of the residual WR erasures in Equation 11 as a function of the position of the item in the test. Figure 2 shows these plots for the first 10 flagged students in the data set at level $\alpha = .001$. For each of these plots, all large residuals were positive. However, no suspicious patterns, for instance, blocks of adjacent items with large positive residuals or other commonalities between flagged students, were found. If they had happened, we could have checked, for instance, whether these patterns were for students with the same teachers or who were close to each other according to the seating plan for the administration of the test.

Robustness Study

As indicated earlier, the assignment of insufficient testing time will lead to much more power of the test because even a few erasures then already point at cheating. However, the price to be paid would be an underestimation of the WR probabilities in Equation 3 for an occasional test taker still able to finish early and review the answers. For this test taker, the critical value in Equation 10 would be smaller than that required by his or her true probabilities. The only way to counter such cases is to make the test generally more conservative by choosing smaller significance levels than intended. A simulation study was conducted to shed some light on the robustness of the procedure with respect to lack of sufficient testing time as well as the size of the adjustments that may be required to offset the effects.

The study was based on the same data set as for the empirical example in the previous section. Basically, the study replicated the analyses in the examples, each time simulating an increase in the number of items the students were unable to reach during their review. To this end, the 65 items in the test were divided into 13 blocks of 5 subsequent items. In each of the 13 runs, the opportunity for review was eliminated for an extra block of items at the end of
FIGURE 2. Plots of the residual WR erasures as a function of the position of the item in the test for the first 10 test takers in the example flagged at $\alpha = .001$. 
the test. The lack of opportunity to review was implemented by coding the absence of an erasure as \( (U_{ni}^{(1)} = 0, U_{ni}^{(2)} = *) \) instead of \( (U_{ni}^{(1)} = 0, U_{ni}^{(2)} = 0) \). This option reflects the fact that when a test taker does not have the time to review the initial response \( U_{ni}^{(1)} = 0 \), we do not know if this person would have confirmed the response or changed it into a correct response. However, when an actual erasure was observed in the data set, the item was still scored as \( (U_{ni}^{(1)} = 0, U_{ni}^{(2)} = 0) \) or \( (U_{ni}^{(1)} = 0, U_{ni}^{(2)} = 1) \), depending on whether the test taker went from a WW answer or to a right answer on the item.

In each run, the final-stage item parameters were reestimated, the null distribution was recalculated for each student, the critical values \( e_n^{*} \) of the test were recalculated for \( \alpha = .05, .01, \) and .001, and the students with significant numbers of erasures were flagged.

Figure 3 shows the percentage of responses in the data sets as a function of the run number in the simulation study. The percentage went down from approximately 40% to 9% for the last run. Remarkably, without any exception, the bayesglm function quickly converged to finite estimates of \( a_{0i} \) and \( b_{0i} \) for each of the items in all data sets.
The changes in the critical value $e_n/C3n$ across the runs for each of the three significance levels used in the study are reported in Figure 4. The plots in this figure are for the first 10 students in the data set. As expected, in each of these plots, the critical values increased with the simulated number of items that the majority of

FIGURE 4. Critical value $e_n/C3n$ as a function of decreasing opportunity of item review for the first 10 students in the data set. In each subsequent run, the opportunity was reduced for the next five items at the end of the tests.

The changes in the critical value $e_n/C3n$ across the runs for each of the three significance levels used in the study are reported in Figure 4. The plots in this figure are for the first 10 students in the data set. As expected, in each of these plots, the critical values increased with the simulated number of items that the majority of
the students was unable to review. For the lack of opportunity to review the last 1 to 10 items or so, the change was relatively small but became much more dramatic for larger numbers of items.

FIGURE 5. Change in the estimates of the estimated probability of a WR erasures on Item 13 as a function of the opportunity of item review for the first 10 students in the data set. In each subsequent run, the opportunity was reduced for the next 5 items at the end of the tests.
Fortunately, the effect of lack of review was only local. As an example, Figure 5 shows the estimated WR probabilities for the first 10 students in the data set on Item 13 in the test during each of the runs in the simulation study. As the item was in the third block and we started reducing the opportunity of review from the end of the test, the effect should not be visible until the 11th run in the simulation study. The changes in the estimates of the WR probabilities reflected this expectation exactly: All estimates remained approximately the same for the item until it was hit by the general lack of review in run 11. The increase in probability shows the level at which we should have estimated the probabilities when these test takers had been able to review these items but the majority of the other test takers had not. The plots also nicely illustrate the earlier claim by Gelman et al. (2008) that their priors still enable us to estimate subsets of parameters with extreme numbers of missing data while not affecting any of the others.

The results in Figure 5 show that, in principle, if the students are able to review their answers up to a certain point in the test, we could still use the procedure in this paper to check the items for any irregular erasures up to this point. However, this partial analysis is only possible when we have dependable information about the point of change.

Concluding Remark

Although the methodology in this paper was suggested by the capacity of optical scanners to detect erased answers on paper tests, it applies equally well to tests delivered by computer. We can then check the log files for answer changes. In fact, one strong additional advantage of this implementation is the possibility of checking the files to see if the items for which no changes were recorded were visited long enough for a second time to interpret the lack of change as the confirmation of the initial response or no review at all.

But for computerized tests, the primary use of the method would be as a check on students secretly consulting with each other about their initial answers. For schools, it is generally impossible to erase answers after their students have logged off. Although, we know of cases where teachers discussed the test with their students at the end of the session and then gave them the right answers to correct their own answers. Such cases would be caught by the method presented in this paper.

References


**Authors**

WIM J. VAN DER LINDEN is Chief Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; wim_vanderlinden@ctb.com. His primary research interests include test theory, applied statistics, and research methods.

MINJEONG JEON is Graduate Student, University of California, Graduate School of Education, Berkeley, CA 94720; mjj@berkeley.edu. Her primary research interests include item response modeling, multilevel modeling, and research methods.

Manuscript received February 24, 2010
Revision received August 30, 2010
Accepted November 6, 2010