

A Paradox in the Study of the Benefits of Test-Item Review

Wim J. van der Linden

CTB/McGraw-Hill

Minjeong Jeon

University of California, Berkeley

Steve Ferrara

CTB/McGraw-Hill

According to a popular belief, test takers should trust their initial instinct and retain their initial responses when they have the opportunity to review test items. More than 80 years of empirical research on item review, however, has contradicted this belief and shown minor but consistently positive score gains for test takers who changed answers they found to be incorrect during review. This study reanalyzed the problem of the benefits of answer changes using item response theory modeling of the probability of an answer change as a function of the test taker's ability level and the properties of items. Our empirical results support the popular belief and reveal substantial losses due to changing initial responses for all ability levels. Both the contradiction of the earlier research and support of the popular belief are explained as a manifestation of Simpson's paradox in statistics.

Background

As a result of the long history of research on item review and answer changing, we know quite a bit: (a) only small numbers of examinees change their initial responses to items and they make changes only on small numbers of items; (b) answer changing appears to benefit most examinees if they change answers for good reasons; (c) answer changing seems to benefit higher ability examinees most often (e.g., Johnston, 1987; McMorris et al., 1991); (d) most answer changes in most studies are wrong-to-right (WR) changes; and (e) the reasons that examinees report for changing initial responses are good reasons.

Benefits and Drawbacks of Answer Review

Many examinees persist in not changing answers (e.g., Skinner, 1987) even though test directions typically encourage examinees to check their work carefully—and even though not changing answers in effect conflicts with the benefits usually accrued by being metacognitive and self-evaluating. Further, evidence from 33 studies on answer changing published between 1928 and 1987 “uniformly [original emphasis] indicates that: (a) the majority of answer changes are from incorrect to correct, (b) most examinees who change their answers improve their test scores” (Benjamin, Clavell, & Shallenberger, 1987, p. 44).

Subsequent studies reached the same general conclusion (e.g., Edwards & Marshall, 1987; McMorris, DeMers, & Schwarz, 1987). While 30 of the 33 studies

in the Benjamin et al. (1987) review were conducted with students in undergraduate, graduate, medical, health care, and U.S. Air Force education programs, results from the three studies of K-12 examinees show similar answer changing behavior and results: answers were changed for approximately 3–7% of items, 40%–60% of those changes corrected previous incorrect answer choices, and changes produced total test score increases far more often than decreases. McMorris et al. (1991) found similar results in a study of 133 Grade 5 and 6 students. A general conclusion from virtually all studies of answer changing is that all examinees typically benefit from reviewing their initial responses and changing answers—if they have reason to doubt that their initial response is the best choice.

Other studies provide refinements to these general conclusions. For example, Skinner (1987) described results from a study of 46 females and 22 males responding to 110 multiple choice items on an introductory psychology examination and cited other studies that indicated that (a) examinees did not change answers, despite recommendations from an examiner that it is to their benefit to do so; and (b) women tend to change answers more often than men. A study of 94 introductory psychology undergraduate students (Johnston, 1987) suggests that higher-achieving students are likely to change answers less frequently and that those changes most often correct initial incorrect responses and produce higher test scores. McMorris et al. (1991) also reported that higher-achieving examinees changed fewer initial responses and had higher rates of WR changes, and that the majority of answer changes for all study subgroups also were from WR.

In a study of Grade 8 readers of different abilities, Casteel (1991) concluded that all readers profited significantly from changing answers on items and that they had a two-to-one chance that the new response would raise rather than lower the final test score. Al-Hamly and Coombe (2005) investigated whether Gulf Arab technology college students should review and change initial answers on subtests of a language proficiency test. On the basis of the answer changes for 286 students, they concluded that examinees should be encouraged to review and change answers. Studies with 125 and 84 introductory psychology undergraduate students demonstrated that cognitive style variables (i.e., field independence/dependence and impulsivity/reflectivity) were not significantly related to canonical structure coefficients, while (positive and negative) effects of answer changes, the number of changes, and examination scores were related to the structure coefficients (Friedman & Cook, 1995). Crocker and Benson (1980) demonstrated that answer changes had no adverse impact on score reliability or item discriminations in a study of Grade 7 examinees on a subtest of a national norm-referenced test.

Reasons for Answer Changes

Still other researchers investigated examinees' reasons for changing initial responses to multiple choice items. For example, McMorris et al. (1987) reported the two most common reasons for changing initial answers, as reported by examinees in graduate measurement classes who were told of the research results on the benefits of changing answers: (a) rethinking the item demands and selecting a better response, and (b) rereading the item and understanding the item response requirements better.

Schwarz, McMorris, and DeMers (1991) found similar results in a follow-up study in five master's level courses in educational measurement.

In another follow-up study, McMorris et al. (1991) surveyed and interviewed 87 Grade 5 and 6 students. The most popular reasons these students gave for changing answers were that they rethought the answer, had guessed on the initial response, learned from a later item, and corrected a clerical error. They also reported with some frequency that they had misread the item at first and remembered something after their initial response. Kruger, Wirtz, and Miller (2005) cited several surveys that indicate that approximately three of four college students believe that answer changing usually lowers scores and that approximately 50% of instructors in one college believe the same. They then conducted four studies on why college students persist in this belief. They found that students experienced more regret (i.e., "'If only . . . ' self-recrimination," p. 9) if they changed a correct to an incorrect answer than if they failed to make the opposite change. In turn, the regret made the experience more memorable and available in memory and led to "counter-factual" beliefs about sticking with initial responses.

Higham and Gerrard (2005) conducted two experiments on the role of metacognition in answer changing behavior on general knowledge tests. In the first experiment, they administered a pretest with deceptive general knowledge questions to one group and nondeceptive questions to a second group. They then required subjects in both groups to change one third of their original responses. This requirement induced a positive experience in the deceptive questions group, as they were able to correct previous incorrect responses, and a negative experience in the other group, as they were forced to make right-to-wrong (RW) answer changes. In the second experiment, the group that experienced the induced negative reaction tended to make fewer answer changes on the posttest. However, the effect appeared only in the metacognition data, in which subjects in the negative experience group did not monitor errors on easy items as carefully as the positive experience group, not in the proportions of correct responses after answer changing.

Edwards and Marshall (1987) surveyed 70 undergraduate students of psychology. Among other findings, students reported that they changed answers on tests because subsequent items provided a clue and that they simply changed their minds (with no explanation). The researchers also reported that one student changed answers if too many responses with the same letter were selected and that students reported changing answers only reluctantly, knowing that sticking with your first impression was the general advice.

Finally, Ferrara et al. (1996) reported a think-aloud study of answer review and changing on a computer-adaptive test. Eighteen of 29 middle school examinees changed answers while retaking a high school graduation test that they previously had failed. Of the 35 reasons for answer changes identified in think-aloud verbal protocols, 17 were to correct calculation errors in the previous response. Other less frequently identified reasons included correcting a misinterpretation of the item and making a more reasonable guess at the correct response. Twelve of the 35 reasons for changing answers were not discernible in the verbal data.

Table 1
Percentages of Answer Changes and Raw-Score Gains Found by McMorris et al. (1987)

Type of Change	End-of-Course Test						Average
	1	2	3	4	5	6	
WR	4.0	3.4	3.7	4.1	2.5	3.8	3.7
WW	1.2	1.0	0.9	1.8	1.0	1.5	1.2
RW	1.8	1.6	1.4	1.6	0.5	1.5	1.5
Gain	2.3	1.7	2.3	2.5	2.0	2.3	2.2
<i>N</i>	21	29	18	21	10	21	

Traditional Method of Analysis

A critical feature of the research tradition above is its reliance on the marginal proportions of answer changes for the group of test takers that is studied. The score gain is directly calculated from these proportions. We illustrate this method by reviewing some of the earlier results published in McMorris et al. (1987).

These authors analyzed the percentages of three different types of answer changes—wrong to right (WR), wrong to another wrong (WW), and right to wrong (RW). Table 1 shows the percentages of changes observed in one part of their study, which used the end-of-course tests in six different master's level classes on educational and/or psychological measurement (McMorris et al., 1987, p. 135). The percentages were calculated as average percentages per student; for instance, the percentage of RW changes for a test was calculated as the average number of RW changes per student divided by the number of items in the test (times 100%). The second to last row of Table 1 contains the net gains in the percentage-correct scores for each of these tests, calculated as the difference between the average percentages of WR and RW changes. All gains were small but positive, varying between 1.7% and 2.5%.

Using data from other tests, the same authors reported slightly increased gains relative to those in Table 1 for classes that were both instructed prior to their test about the beneficial effects of item review and encouraged to change initial answers believed to be wrong during their review (McMorris et al., 1987, p. 137).

The authors also found that answers to items with lower *p*-values tended to be changed more frequently than answers to items with higher *p*-values ($r = -.55$) but with relatively more WW changes as a result. Further, the number of changes correlated positively with the classical *D*-index for item discrimination ($r = .27$; McMorris et al., 1987, p. 139).

Modeling Answer Changes

In our research, we took a new look at the problem of the benefits of item review and used an item response theory (IRT) approach in which the probabilities of all possible answer changes were modeled as a function of the test taker's ability and the properties of the items. The importance of the role of the item properties is illustrated by the earlier correlations between the answer changes and classical item indices reported by McMorris et al. (1987). The dependence of the probabilities on the test

taker's ability is somewhat more complex: We expect more able test takers to be less likely to make an incorrect initial answer. But if they make one, they are expected to be more likely to change it into a correct answer during item review than less able test takers. An analogous argument is assumed to hold for the probabilities of correct initial answers and subsequent decisions about changes.

This type of modeling previously has been used to detect numbers and patterns of WR changes on answer sheets, found by optical scanners, that may be indicative of cheating (van der Linden & Jeon, in press). This same modeling extends well to the study of answer changes during item review. In addition, it enables derivation of an expression for the expected benefits of answer changes using simple probability calculus. We explain this type of modeling and then show how to derive a measure of expected benefits from the model.

Two-Stage Answering Process

In order to make the fine distinctions among the probabilities of initial answers and answer changes, we model item review explicitly as a two-stage process. In the first stage, the test taker produces an initial answer on an item; in the second, the test taker reviews the initial answer and either confirms or changes it.

Both paper-and-pencil and computerized tests allow us to record the two types of answers separately. For paper tests, modern optical scanners can be set to detect both erased initial answers and new answers on the answer sheets. If no erasure is observed, and the test taker had enough time for review and can be assumed to be interested in maximizing the test score, it is safe to conclude that the initial answer was confirmed. For computerized testing, the recording of the same information is straightforward; all keystrokes are logged automatically, and the logfiles can be checked to see when test takers went back to review an item and, while doing so, whether they changed or confirmed the initial answer.

Differences between initial and final answers may be the result of different scenarios. As noted earlier, students state different reasons for revising an initial answer. They may do so, for instance, because they (a) rethought it and came to a different conclusion or discovered a clerical error during a second visit to the item, (b) guessed quickly during their first attempt and returned at the end of the test for a more serious attempt or just to correct a recording error, or (c) later realized that they had misread the item or remembered something else about their initial answer. In this study, however, the focus is on the change in the correctness of the first and final responses, not on the nature of the possible scenario that underlies the change. In principle, the latter does not imply anything about the correctness of the first and final answers: rethinking an answer does not necessarily mean that it was wrong and will be corrected, a discovered "clerical error" actually may have been correct, a later brain wave is not necessarily better than an initial wave, and so on. In fact, it is precisely the purpose of this study to find out if the net effect of such scenarios tends to be a gain or a loss.

Models for the Two Stages

A correct initial answer for test taker n on item i will be denoted as $U_{ni}^{(1)} = 1$, whereas $U_{ni}^{(1)} = 0$ will be used to denote an incorrect initial answer. Likewise, $U_{ni}^{(2)} = 1$ and $U_{ni}^{(2)} = 0$ will be used to denote the correctness of the final answer.

For initial answers, a regular response model is assumed to hold. The assumption is always met for testing programs that use such models to calibrate their pretested items and check their fit. The case of a program with dichotomous items and the three-parameter logistic (3PL) model as its operational response model is considered; the modifications required for other types of response models will be discussed later.

The 3PL model gives the probability of a correct answer for test takers $n = 1, \dots, N$ with ability $\theta_n \in R$ on items $i = 1, \dots, I$ as

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

where $b_i \in R$ can be interpreted as the difficulty parameters of item i , $a_i \in (0, \infty]$ as a parameter for its discriminating power, and $c_i \in [0, 1]$ represents the height of the lower asymptote required to deal with the effects of guessing. Throughout this paper, it is assumed that these parameters have been estimated with enough precision to treat their values as known.

The initial answers are likely to have an impact on the final answers. Hence, the usual assumption of local independence is omitted, and we specify the probabilities of success for the final answers conditionally on the initial answers. That is, for $U_{ni}^{(1)} = 0$ and $U_{ni}^{(1)} = 1$, we consider two different response probabilities for $U_{ni}^{(2)} = 1$, namely the probabilities $\Pr\{U_{ni}^{(2)} = 1 \mid U_{ni}^{(1)} = 0\}$ and $\Pr\{U_{ni}^{(2)} = 1 \mid U_{ni}^{(1)} = 1\}$. Once these conditional probabilities are known, the probabilities of $U_{ni}^{(2)} = 0$ given the two possible initial responses also are known.

For the case of the 3PL model fitting the initial answers, it makes sense to model the conditional probabilities $\Pr\{U_{ni}^{(2)} = 1 \mid U_{ni}^{(1)} = 0\}$ and $\Pr\{U_{ni}^{(2)} = 1 \mid U_{ni}^{(1)} = 1\}$ as versions of the same initial model with a different treatment of its parameters. The following modifications are made: (i) item parameters a_i and b_i are treated as free parameters; (ii) the c_i parameters are set equal to zero; and (iii) the θ_n parameters are set equal to their initial values, that is, $\theta_n = \theta_n^{(1)}$.

The modifications are motivated as follows. The only thing that counts is a correct representation of the changes in the response probabilities for the items as a result of the review. Therefore, first, as initially correct and incorrect answers may have a different impact on the probability of a final answer, we need free parameters a_i and b_i in the two final-stage models to allow for the possibility of the different changes. Second, the 3PL model is based on the assumption of knowledge or random guessing. If the initial answer was guessed but the test takers now believe that they know the correct answer, they will choose it. If they still do not know the answer, there is no reason to guess again. Therefore, for both final-stage models, we set $c_i = 0$. The models thus effectively become 2PL models. Of course, the assumption of knowledge or random guessing itself may be wrong, but this is why we always should check the fit of the models for the two stages. Third, for the 3PL model in (1), only the difference between the θ_n and b_i parameters is identified; a change of the former introduces the same change in the response probability as the opposite change of the latter. However, both the a_i and b_i parameters already are free to record such changes at the level of the individual items. We therefore fix the θ_n parameters at their

initial values. As shown in the next section, the last assumption leads to a convenient statistical treatment of the two models.

Formally, the two conditional final-stage models can be written as

$$\Pr \{U_{ni}^{(2)} = 1 \mid U_{ni}^{(1)} = 0\} = \frac{\exp [a_{0i}(\theta_n^{(1)} - b_{0i})]}{1 + \exp [a_{0i}(\theta_n^{(1)} - b_{0i})]}, \quad (2)$$

$$\Pr \{U_{ni}^{(2)} = 1 \mid U_{ni}^{(1)} = 1\} = \frac{\exp [a_{1i}(\theta_n^{(1)} - b_{1i})]}{1 + \exp [a_{1i}(\theta_n^{(1)} - b_{1i})]}, \quad (3)$$

with (a_{0i}, b_{0i}) and (a_{1i}, b_{1i}) parameters to be estimated from the final responses.

Parameter Estimation and Model Fit

The two item parameters in the conditional response models in (2) and (3) are estimated from different sets of data; the former is estimated from the incomplete data matrix with the responses $U_{ni}^{(2)} = u_{ni}^{(2)}$ for test takers with an initial response $U_{ni}^{(1)} = 0$, and the latter is estimated from the incomplete matrix with $U_{ni}^{(2)} = u_{ni}^{(2)}$ for the test takers with $U_{ni}^{(1)} = 1$. The missing data in both complementary matrices are missing at random; hence, their absence does not bias the estimates in any way. Also, the loss of efficiency due to missing data is not as serious as for the regular 3PL model because we have to estimate only $2I$ instead of $3I + N$ parameters.

Assuming the $\theta_n^{(1)}$ estimates are accurate enough, the estimation problem boils down to the estimation of the intercept and slope parameters in simultaneous logistic regressions of the responses $U_{ni}^{(2)}$ on the first-stage ability estimates $\hat{\theta}_n^{(1)}$ for the items. In the empirical study below, we used a Bayesian approach to logistic regression with the priors proposed in Gelman, Jakulin, Pittau, and Su (2008). These priors are independent Cauchy distributions with location zero and scale equal to 10 and 2.5 for the intercept and slope parameters, respectively. Use of the priors saves logistic regression from possible problems due to sparse data such as quasi-complete separation or even complete separation while not adding any significant subjective information to the estimates. In fact, they can be shown to be less informative than the observation of one half of a correct and one half of an incorrect response for the logistic probability with intercept equal to one and slope equal to zero (Gelman et al., 2008). The procedure is available as the function `bayesglm` in the `arm` package for multilevel analysis in *R*, developed by these authors. For more details, we refer to the publication by Gelman et al. (2008).

Methods for checking the fit of a logistic regression model are amply available in the literature on logistic regression. In the empirical study below we used the Gelman and Pardoe (2006) R^2 as a measure for the explained variance in the responses.

Opportunity to Review

The assumption of test takers reviewing all items during the final stage can be met only when they have enough time to do so and are interested in maximizing

their scores. The models in (2)–(3) therefore should not be used for tests with time limits that do not allow for item review or with test takers unmotivated to do so. For such cases, however, interest in the benefits of item review among test takers seems unlikely.

The effects of violations of the assumption of full review are known, though. They lead to two different types of potential confounding. First, for the case of an item with an incorrect answer that was not changed, it is no longer clear whether the test taker confirmed it or was unable to review it because of lack of time. Consequently, it is unknown whether the answer vector for the item should be equal to $(U_{ni}^{(1)} = 0, U_{ni}^{(2)} = 0)$ or $(U_{ni}^{(1)} = 0, U_{ni}^{(2)} = *)$, with $*$ denoting a missing response. Second, for the case of an item with a correct answer that was left unchanged, it is unclear whether its vector should be equal to $(U_{ni}^{(1)} = 1, U_{ni}^{(2)} = 1)$ or $(U_{ni}^{(1)} = 1, U_{ni}^{(2)} = *)$.

For the first case, if we used $(U_{ni}^{(1)} = 0, U_{ni}^{(2)} = 0)$ instead of $(U_{ni}^{(1)} = 0, U_{ni}^{(2)} = *)$, the result would be an underestimation of the success probability in (2). Likewise, for the second case, the use of the wrong response vector would lead to an overestimation of the success probability in (3).

Fortunately, however, as the benefits measure introduced below is a weighted sum of both probabilities, the two biases tend to counterbalance. We will make this conclusion more specific once its expression has been derived.

Other Response Models

The same approach can be applied for any operational response model currently in use for a testing program. For other dichotomous models, the modification is straightforward. For polytomous models such as the partial-credit model, the final-stage models are of the same type but with a different conditional version given each possible initial response. As these models typically do not involve any guessing parameters, all item parameters should be free but the ability parameters should be fixed at their initial values.

Expected Benefits

Using the first-stage and second-stage models, it is possible to derive an expression for the expected benefits due to the review of an item as a function of the test taker's ability. The derivation is as follows: If there is no review, the expected score of a test taker with ability θ on item i is equal to

$$\begin{aligned} \mathcal{E}(U_i^{(1)}|\theta) &= 1 \cdot P(U_i^{(1)} = 1|\theta) + 0 \cdot P(U_i^{(1)} = 0|\theta) \\ &= P(U_i^{(1)} = 1|\theta), \end{aligned} \tag{4}$$

where $P(U_i^{(1)} = 1|\theta)$ is the success probability for the initial answers given by the 3PL model in (1).

However, upon review, the test taker’s expected score is

$$\begin{aligned}
 \mathcal{E}(U_i^{(2)}|\theta) &= 1 \cdot [P(U_i^{(2)} = 1|U_i^{(1)} = 1, \theta)P(U_i^{(1)} = 1|\theta) \\
 &\quad + P(U_i^{(2)} = 1|U_i^{(1)} = 0, \theta)P(U_i^{(1)} = 0|\theta)] \\
 &+ 0 \cdot [P(U_i^{(2)} = 0|U_i^{(1)} = 1, \theta)P(U_i^{(1)} = 1|\theta) \\
 &\quad + P(U_i^{(2)} = 0|U_i^{(1)} = 0, \theta)P(U_i^{(1)} = 0|\theta)] \\
 &= P(U_i^{(2)} = 1|U_i^{(1)} = 1, \theta)P(U_i^{(1)} = 1|\theta) \\
 &\quad + P(U_i^{(2)} = 1|U_i^{(1)} = 0, \theta)P(U_i^{(1)} = 0|\theta).
 \end{aligned}
 \tag{5}$$

The expression in (5) writes the expected score as the result of the compound event of a correct initial answer confirmed during the review or an incorrect initial answer replaced by a correct one.

Subtracting (4) from (5) gives the expected benefit due to the review of item i as a function of θ :

$$\begin{aligned}
 \beta_i(\theta) &= \mathcal{E}(U_i^{(2)} - U_i^{(1)}|\theta) \\
 &= P(U_i^{(2)} = 1|U_i^{(1)} = 1, \theta)P(U_i^{(1)} = 1|\theta) \\
 &\quad + P(U_i^{(2)} = 1|U_i^{(1)} = 0, \theta)P(U_i^{(1)} = 0|\theta) - P(U_i^{(1)} = 1|\theta).
 \end{aligned}
 \tag{6}$$

The expected benefit for the number-correct score on the test as a function of θ is the sum of the benefits in (6) across all items; that is,

$$\beta(\theta) = \sum_{i=1}^N \beta_i(\theta).
 \tag{7}$$

All probabilities necessary to calculate the expected total-score benefits $\beta(\theta)$ are given in (1), (2), and (3). Once their item parameters are known, $\beta(\theta)$ can be calculated immediately.

When a test taker does not have the opportunity to review an item, the first term of (5) will be overestimated but the second term underestimated. The net effect thus depends on the size of these two biases (i.e., how often the violations occur) as well as the probability of success for the initial answer. The latter suggests no effect on the estimated benefits of item review in the middle of the ability range (i.e., when $P(U_{ni}^{(1)} = 1 | \theta)$ is close to .5) but underestimation for low and overestimation for high abilities. However, for lowest and highest abilities, the response curves in (2) and (3) are close to 0 and 1, respectively, and for those cases the bias generally will be negligible.

Empirical Study

The models in (2)–(3) were fitted and the expected item-level and total-score benefits in (6)–(7) calculated with the 3PL model for a data set from a large-scale

Table 2
Total Numbers of Changes Across All Items and Students

Initial Answer	Final Answer	
	0	1
0	56,587	11,543
1	1,454	96,481

assessment program. The data were for a 65-item mathematics test administered to 2,555 Grade 3 students. All items had shown excellent fit to the 3PL model during pretesting. Their final calibration was based on marginal maximum-likelihood estimation with $\theta \sim N(0, 1)$. The test was administered under untimed conditions. In addition, the test administrators were required to read scripted directions that encouraged the students to read the instructions and questions carefully. Upon completion of the tests, the students were encouraged to review their answers before turning in their test materials. As for possible motivation problems, all students were Grade 3 students who generally are known to take assignments with tests seriously.

The answer sheets of the students were scanned to record their initial and final responses. The initial responses, $U_{ni}^{(1)}$, were used to estimate the students' ability parameters, $\theta_n^{(1)}$. The estimation method was expected a posteriori (EAP) estimation with a uniform prior over $[-4, 4]$.

The set of final responses, $U_{ni}^{(2)}$, was split into the two subsets for $U_{ni}^{(1)} = 0$ and $U_{ni}^{(1)} = 1$. The subsets were used to estimate the item parameters in the models in (2) and (3), respectively using the `bayesglm` function in the `arm` package for *R* with the prior distributions for the slope and intercept parameters described earlier. The package assumes the logistic regression parameterization $a_i^* \theta + b_i^*$. After it was used, the estimated intercept parameters were transformed back to the scale of the regular b_i parameters using $b_i = -a_i^* b_i^*$.

The fit of the models was checked using Gelman and Pardoe's (2006) R^2 . The result was $R^2 = .487$ and $R^2 = .906$ for (2) and (3), respectively, which means that 48.7% and 90.6% of the variance of the responses was explained by these models. The percentages imply correlations between predicted and observed final responses equal to approximately .70 and an extremely high value of .95, respectively. These percentages and correlations point at sufficient fit of both models to analyze trends across items. We are unable to offer an explanation of the differences between these results for the two models.

Results

Table 2 gives the total numbers of answer changes across all students and items in our data set. For the WR and RW changes, the numbers were equal to 11,543 and 1,454, respectively. Thus, the number of changes from WR was approximately 8.5 times larger than the number from RW. For comparison with the percentages of changes in the earlier study by McMorris et al. (1987), we also calculated the percentages of WR and RW changes in the data set; they were equal to 7% and

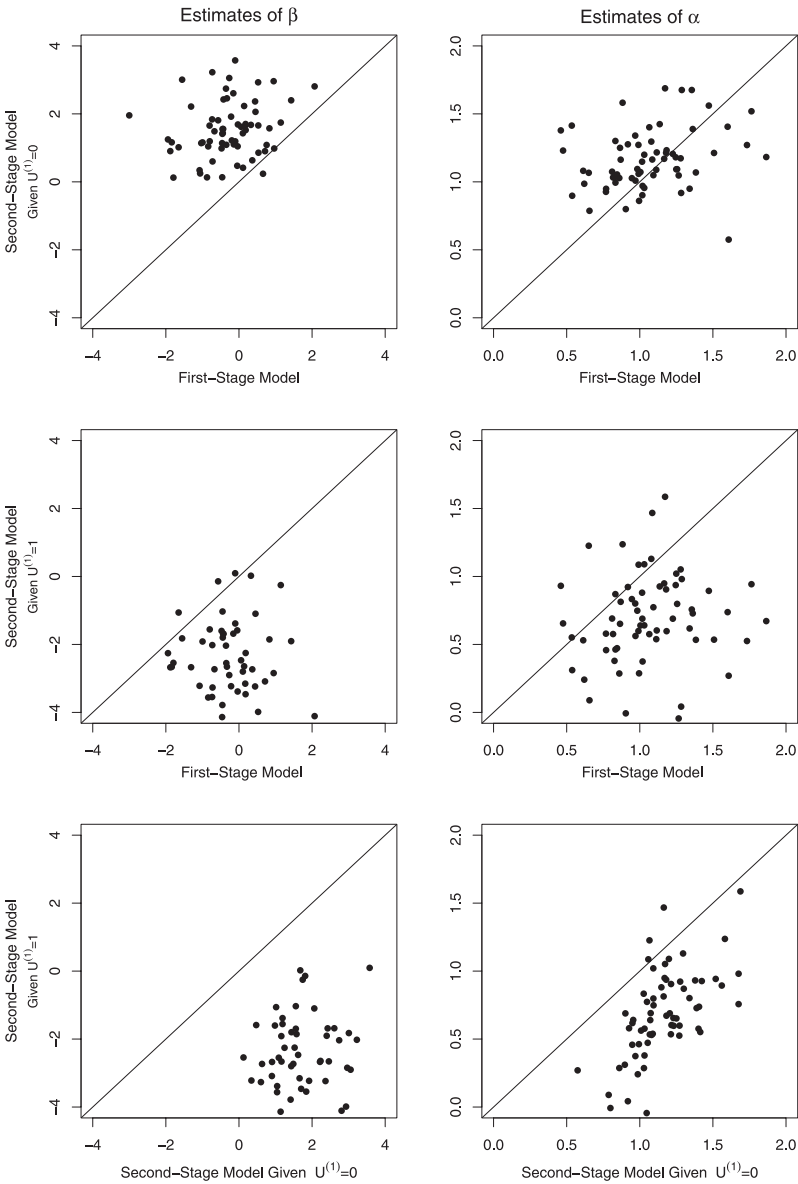


Figure 1. Pairwise plots with the estimates of the b_i and a_i parameters in the first-stage and two second-stage models against each other.

0.8%, respectively. The difference in percentage was thus $7\% - 0.8\% = 6.2\%$ —a result that seems to point at a much more convincing gain for item review than the maximum difference of 2.5% found by these earlier authors (Table 1).

The differences between the estimated a_i and b_i parameters in the three response models are summarized in Figure 1. The following conclusions can be drawn from the plots. First, the difficulty parameters b_i in the second-stage model for the case

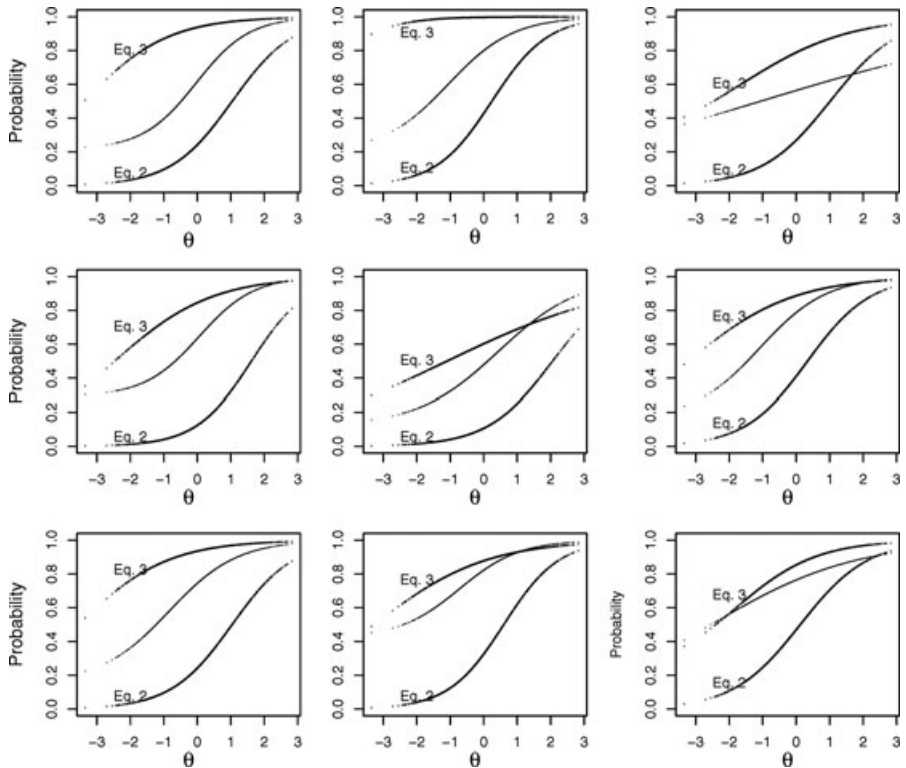


Figure 2. Plots of the response functions for the initial-stage model and the second-stage models in Equations 2 and 3 for the first nine items in the test.

of an incorrect initial answer tend to be higher than those in the first-stage model. Apparently, it is easier to give a correct initial answer than change to the correct answer during review. Second, the opposite is observed for the difficulty parameters in the second-stage model for the case of a correct initial answer; these tend to be lower than the difficulty parameters in the first-stage model. It thus appears easier to maintain a correct initial answer during review than to give one directly. Third, the discrimination parameters a_i for the second-stage model for the case of a correct initial answer generally are lower than the initial discrimination parameters. This result implies that our second conclusion is somewhat less critical than it may seem; the probability of maintaining a correct initial answer increases relatively slowly with the ability of the test taker. Fourth, the opposite holds for the probability of changing an incorrect initial answer into a correct one; this probability increases more sharply with ability (in the neighborhood of the difficulty parameter). Fifth, the final two plots in Figure 1 summarize the previous differences between the parameter estimates for the two second-stage models. Both the b_i and a_i parameters for the second-stage model given an incorrect initial answer are considerably higher than for the model given a correct initial answer.

Figure 2 shows the estimated response functions for the three models for the first nine items in the test. These functions illustrate the conclusions from Figure 1

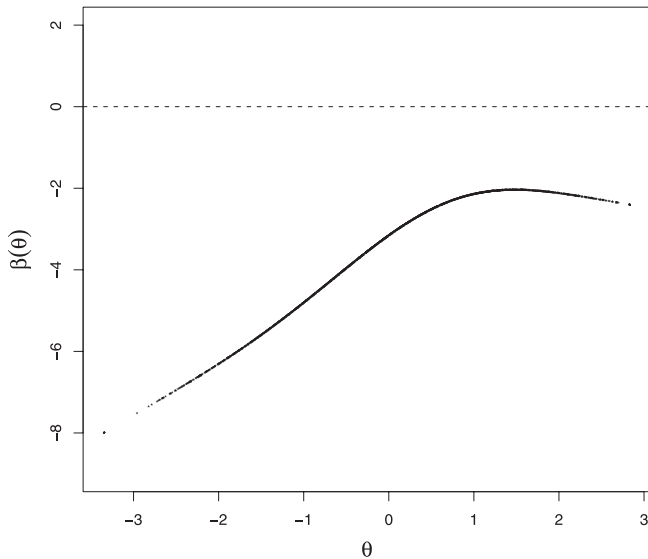


Figure 3. Expected total-score benefits $\beta(\theta)$ plotted as a function of θ .

graphically. The three response functions for the first item (top left plot) are typical. The function in the middle is for the initial response. The one to the right of it is for the final response given an incorrect initial response. The one to the left, with a lower difficulty as well as a relatively less steep slope, is for the case of a correct initial response.

The main result from this research is given in Figure 3, which shows the expected total-score benefits as a function of the test taker's ability, $\beta(\theta)$. The benefits are negative for all ability values, with considerable losses for the lower abilities. At $\theta = -2$, the loss is even greater than 10% of the range of possible number-correct scores on this 65-item test. The lowest loss occurs close to $\theta = 1.3$, but even then the penalty for item review amounts to more than two score points. The plots in Figure 4 illustrate the expected item-level benefits for the first nine items in the test. The general trend clearly is negative. The loss at the lower ability levels is uniform across all items, but positive gains did occur for an occasional item at some of the higher ability levels. However, these gains were never larger than 0.1 score point.

Our findings clearly contradict conclusions offered in the long tradition of empirical research on the benefits of answer changing. In fact, they even seem inconsistent with the strong gains we were able to report in Table 2. How is this possible? A computational error? Serious bias in some of the parameter estimates? Any other deficiency?

Simpson's Paradox

The answer may lie in a manifestation of a phenomenon more generally known as Simpson's paradox in statistics (Simpson, 1951). The paradox may occur when the relationship between two variables is further specified by conditioning on a third variable. A necessary condition for it to occur is the third variable correlating with

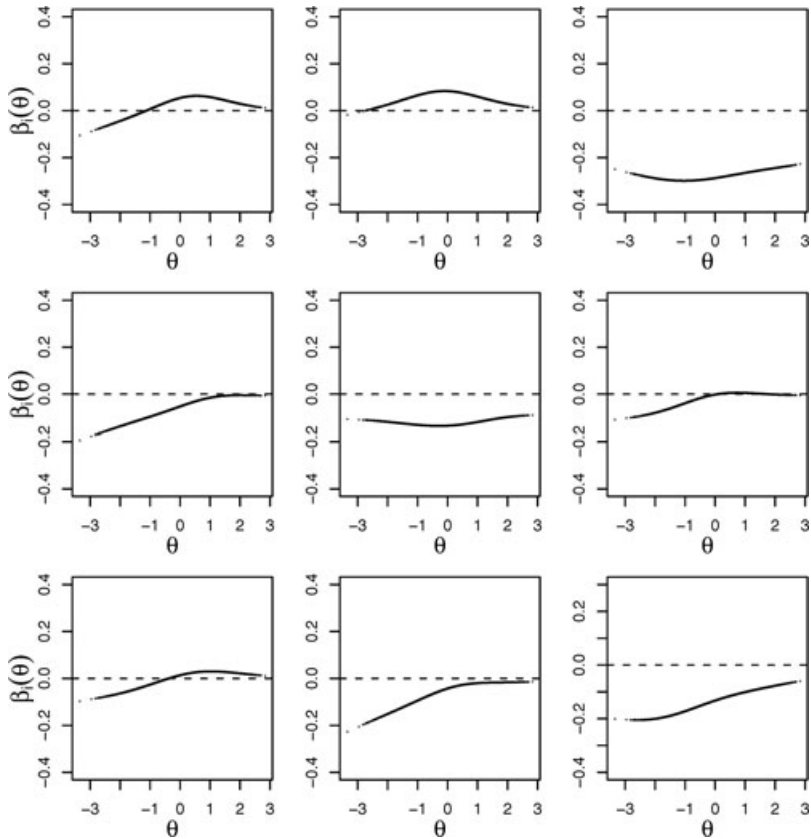


Figure 4. Expected item-level benefits $\beta_i(\theta)$ for the first nine items in the test plotted as a function of θ .

both initial variables. If the condition is met, the relationship may change or even completely reverse. A popular introduction to the paradox is given in Wagner (1982).

Although the phenomenon is referred to as Simpson's paradox in table analysis, it is known in other areas by such names as "spurious correlation" or "unbalancedness" of the research design. The methodological literature generally refers to the third variable as a confounding variable. Simpson's paradox may repeat itself when conditioning on more than one variable. A table illustrating both the reversal of a relationship between two variables conditional on a third variable and an opposite reversal conditional on the combination of the third and a fourth variable can be found in Novick (1983). The reverse phenomenon (i.e., change of the relationship when collapsing a two-way table over a third variable) is a well-known issue in table analysis as well (e.g., Fienberg, 1977, sect. 3.8).

A clear example of Simpson's paradox occurred in a study of sex bias in graduate admission at the University of California at Berkeley. Bickel, Hammel, and O'Connell (1975) reported that bias against female applicants observed in the total set of admission data disappeared, and occasionally even reversed, when the data were analyzed separately for each graduate program. The confounding variable in

this study was the severity of the admission procedure; females tended to apply more frequently to departments with higher admission thresholds.

Other examples are found in the medical literature. One of the better-known examples is the comparison of the success rates in removing kidney stones between open and minimally invasive surgery in Charig, Webb, Payne, and Wickham (1986). The data set later was discovered to have a rather dramatic example of the paradox (Julious & Mullee, 1994): minimally invasive surgery generally was more successful than open surgery, but when the successes of the two alternatives were analyzed separately for small (<2 cm) and large (≥ 2 cm) stones, the success rates for *both* categories reversed. The reason for the reversal was that the size of the kidney stone correlated both with the choice of operation technique and the success rates.

It is not necessary for the third variable to be observable; the paradox may occur when conditioning on a latent variable. As a matter of fact, this is exactly what happened when we added the ability variable to our analysis in Table 2. Table 2 displays only the marginal association between the initial and final response for all students. However, as shown in Figure 2, ability correlates with both responses. Consequently, the relationship between the initial and final responses was able to reverse when we conditioned on ability.

Simpson's paradox is not a formal paradox. It creates the illusion of a paradox because of erroneous interpretation of the data. In our case, the illusion arises when we interpret the proportion of answer changes for the *population* in Table 2 as probabilities valid for each *individual* student. However, the students with WR changes in the upper right cell are not the same as those with RW changes in the lower left cell. Ending up in the former requires an incorrect initial response; the latter requires a correct response. As reflected by the 3PL model for the probability of an initial response, the probabilities of these two responses are high at the opposite ends of the ability scale. Thus, there exists no single probability of a WR or RW change for all test takers; in order to know the actual probabilities, we have to condition on their ability level.

Figure 5 illustrates the "paradox" graphically. It shows the three response curves for the middle item in the second row of Figure 2, this time with the left-hand side curve replaced by its complement [one minus the probability in (3)]. As a result, the two second-stage curves now represent the conditional probabilities of an RW and WR change—two of the key probabilities that drive the expected benefits of item review in (7). Observe that the curve for the probability of an RW change runs higher over most of the displayed part of the ability scale than the one for a WR change. Their order thus is the reverse of what was suggested by Table 2, hence the negative expected benefits for the fifth item in Figure 4. A similar trend across all items explains the negative total-score benefits in Figure 3.

As an aside, note also that the two probabilities of RW and WR changes are highest at the low and high end of the ability scale, respectively; this is exactly where the probabilities of the initial answers required for the two changes are low. This observation explains why percentages of answer changes as in Tables 1 and 2 tend to be low.

How is it possible that the popular belief among test takers that one generally should be inclined to retain the initial answers is in agreement with the more complicated IRT modeling in this paper rather than the simple analysis in Tables 1 and 2? The explanation is perceptual. Test takers never see any data as in Table 1 or

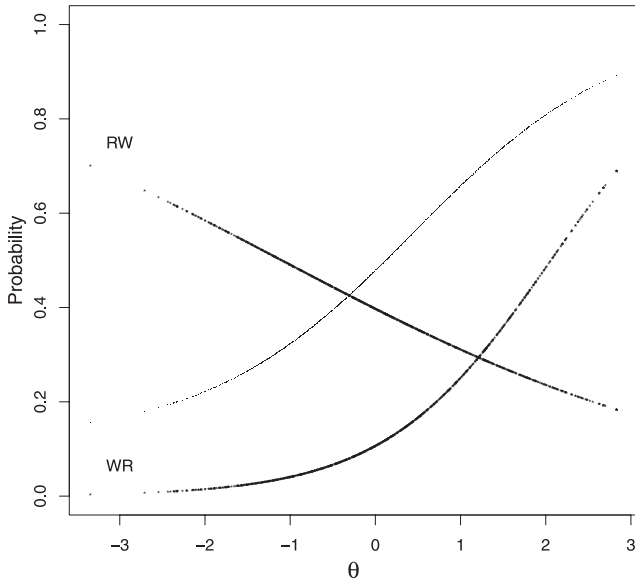


Figure 5. Plot of the three response functions for the fifth item in the test with the curve for the second-stage model in Equation 3 replaced by its complement. The two second-stage curves now represent the probabilities of a RW and WR change.

2; they are thus automatically prevented from the fallacy of interpreting group-based proportions as individual probabilities. But they do have lifelong experience with the struggle between initial answers and second guesses at their own level of ability. And, apparently, they may be governed by aversion to the potential “If only . . . self-recrimination” reported by Kruger et al. (2005). This conditional experience is exactly what is modeled in Figure 3.

We are aware of one other case of Simpson’s paradox as a perceptual explanation of a popular belief: the “hot hand” in basketball. According to this belief, shooters have a higher chance of a hit after a previous hit than after a miss. Tversky and Gilovich (1989) tested the belief by subjecting free-throw data for nine National Basketball Association players to rigorous statistical hypothesis testing. Their results led them to reject the hot hand belief in favor of the alternative (that the probabilities of successful free throws are independent). The point to note, however, is that they tested the belief separately for the data for each individual player. Wardrop (1995) reanalyzed the same data aggregated across all nine players and obtained a table suggesting support for the hot hand. The confounding factor that explains these contradictory results is the variation in free-throw-shooting skill between the players. When spectators follow a game, they observe streaks of hits by the more-skilled players alternated with misses (and occasional successes) by the less-skilled players. If they followed just one player during the game, they would observe independent shots with a success probability determined by the player’s skill.

The papers by Tversky and Gilovich (1989) and Wardrop (1995) are interesting for another reason as well. Marginal dependence between successes across basketball

players but conditional independence given a player's skill is exactly the same as positive dependence between responses to different test items across test takers but local independence given their ability—a standard assumption in IRT. The difference between these two types of independence might very well be the oldest manifestation of Simpson's paradox known to exist in the history of test theory.

Discussion and Conclusions

In this study, we examined the question, “which belief about answer changing in multiple choice items is correct: The popular belief that examinees should go with first responses, and not change them? Or the expert belief, based on 80 years of empirical studies, that examinees improve their raw scores by reviewing and changing initial responses?” Our empirical results support the popular belief about answer changing. And we demonstrated that the expert belief may have been fed by inferential errors: (a) ignoring the probabilities of the initial answers required for a change; and (b) interpreting proportions of answer changes across all examinees as if they were probabilities that applied to each individual examinee thereby disregarding the differences between their abilities. These errors create the conditions necessary for a phenomenon known in statistical literature as Simpson's paradox.

Why are these issues important? Because they are relevant to helping examinees optimize their test scores on achievement tests. More specifically, they are relevant to (1) scripting test administration directions to examinees, (2) the advice that teachers give to their students as they prepare for state and other assessments, (3) advice that test preparation companies convey to their high-paying customers, and (4) the debate over item review in computerized adaptive test administrations. Our findings seem to favor recommending that test takers make few cautious answer changes only in cases where they detect a clear calculation error or an entirely overlooked datum.

Further, these findings add a new consideration to the still unconcluded debate on whether to allow item review and answer changing in computerized adaptive testing (CAT; e.g., Wise, 1996). This debate stems primarily from attempting to make CAT administrations mimic aspects of paper-and-pencil test administrations and the results from both interchangeable. A concern from the CAT side is that answer changes lead to post hoc adjustment of ability estimates and therefore less than optimal adaptation with a penalty in the form of larger standard errors. In addition, when offered the opportunity of item review, test-wise examinees may believe that they can “game” the test by responding incorrectly to early items to get subsequent easier items and then return to the earlier items to correct their answers (which leads to larger standard errors as well). (See Vispoel, Clough, & Bleiler, 2005 for an evaluation of six strategies for using item difficulty to change answers on CATs.) Our results suggest that it may not be necessary to provide an opportunity to review and change answers to previous items in CAT because little may be gained and much risked.

As already noted, Simpson's paradox may change again when conditioning on an additional variable. In principle, it thus is possible that a new variable may be found that modifies the relationship between the benefits of answer changes (e.g., a personality variable) and the test takers' ability. However, a necessary condition

for such variables to have any impact on the probabilities of the answer changes in the benefits measure in (6) is violation of the unidimensionality assumption in the response models adopted in this study. Thus, evaluation of model fit also is a good criterion for the success of a search for a new relevant variable.

This study is unlikely, by itself, to settle the debate about item review in CAT and paper-and-pencil testing. The study was based only on the responses to one 65-item Grade 3 mathematics test. We hope that other researchers use either the modeling described here or other methods that account for examinee ability to conduct independent replications for other grade levels, content areas, and areas outside of achievement testing. We also encourage researchers to further investigate how examinees think when they review (and consider changing) initial answers in order to extend the existing knowledge beyond the limited findings in the studies we cited in the Background section.

References

- Al-Hamly, M., & Coombe, C. (2005). To change or not to change: Investigating the value of MCQ answer changing for Gulf Arab students. *Language Testing*, 22, 509–531.
- Benjamin, L. T., Cavell, T. A., & Shallenberger, W. R. (1987). Staying with initial answers on objective tests: Is it a myth? In M. E. Ware & R. J. Millard (Eds.), *Handbook on student development: Advising, career development, and field placement* (pp. 45–53). Hillsdale, NJ: Lawrence Erlbaum.
- Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admission: Data from Berkeley. *Science*, 187, 398–404.
- Casteel, C. A. (1991). Answer changing on multiple-choice tests among eighth-grade readers. *Journal of Experimental Education*, 59, 300–309.
- Charig, C. R., Webb, D. R., Payne, S. R., & Wickham, J. E. A. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous necrolithotomy, and extracorporeal lithotripsy. *British Medical Journal*, 292, 879–882.
- Crocker, L., & Benson, J. (1980). Does answer changing affect test quality? *Measurement and Evaluation in Guidance*, 12, 233–239.
- Edwards, K. A., & Marshall, C. (1987). First impressions on tests: Some new findings. In M. E. Ware & R. J. Millard (Eds.), *Handbook on student development: Advising, career development, and field placement* (pp. 58–60). Hillsdale, NJ: Lawrence Erlbaum.
- Ferrara, S., Albert, F., Gilmartin, D., Knott, T., Michaels, H., Pollack, J., Schuder, T., Vaeth, R., & Wise, S. L. (1996, April). *A qualitative study of the information examinees consider during item review on a computer-adaptive test*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.
- Fienberg, S. E. (1977). *The analysis of cross-classified categorical data*. Cambridge, MA: MIT Press.
- Friedman, S. J., & Cook, G. L. (1995). Is an examinee's cognitive style related to the impact of answer changing on multiple-choice tests? *Journal of Experimental Education*, 63, 199–213.
- Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2, 1360–1383.
- Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48, 241–251.

- Higham, P. A., & Gerrard, C. (2005). Not all errors are created equal: Metacognition and changing answers on multiple-choice tests. *Canadian Journal of Experimental Psychology*, 59, 28–34.
- Johnston, J. J. (1987). Answer-changing behavior and grades. In M. E. Ware & R. J. Millard (Eds.), *Handbook on student development: Advising, career development, and field placement* (pp. 53–54). Hillsdale, NJ: Lawrence Erlbaum.
- Julious, S. A., & Mullee, M. A. (1994). Confounding and Simpson's paradox. *British Medical Journal*, 309, 1480–1481.
- Kruger, J., Wirtz, D., & Miller, D. T. (2005). Counterfactual thinking and the first instinct fallacy. *Journal of Personality and Social Psychology*, 88, 725–735.
- McMorris, R. F., DeMers, L. P., & Schwarz, S. P. (1987). Attitudes, behaviors, and reasons for changing responses following answer-changing instruction. *Journal of Educational Measurement*, 24, 131–143.
- McMorris, R. F., Schwarz, S. P., Richlichi, R. V., Fischer, M., Buczek, N. M., Chevalier, C. L., & Meland, K. A. (1991). *Why do young students change answers on tests?* (ERIC Document Reproduction Service, ED 342 803).
- Novick, M. R. (1983). The centrality of Lord's paradox and exchangeability for all statistical inference. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord* (pp. 41–53). Hillsdale, NJ: Erlbaum.
- Schwarz, S. P., McMorris, R. F., & DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. *Journal of Educational Measurement*, 28, 163–171.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13, 238–241.
- Skinner, N. F. (1987). Switching answers on multiple-choice questions: Shrewdness or shibboleth? In M. E. Ware & R. J. Millard (Eds.), *Handbook on student development: Advising, career development, and field placement* (pp. 54–55). Hillsdale, NJ: Lawrence Erlbaum.
- Tversky, A., & Gilovich, T. (1989). The cold facts about the "hot hand" in basketball. *Chance*, 2, 16–21.
- van der Linden, W. J., & Jeon, M. (in press). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics*. doi: 10.3102.1076998610396899.
- Vispoel, W. P., Clough, S. J., & Bleiler, T. (2005). A closer look at using judgments of item difficulty to change answers on computer adaptive tests. *Journal of Educational Measurement*, 42, 331–350.
- Wagner, H. (1982). Simpson's paradox in real life. *The American Statistician*, 36, 46–48.
- Wardrop, R. L. (1995). Simpson's paradox and the hot hand in basketball. *The American Statistician*, 49, 24–28.
- Wise, S. L. (1996, April). *A critical analysis of the arguments for and against item review in computer adaptive testing*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

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 a 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.

STEVE FERRARA is Principal Research Scientist, CTB/McGraw-Hill, 1200 G St., NW, Suite 1000, Washington, DC 20005; steve_ferrara@ctb.com. His primary research interests include assessment design, examinee cognitive processing, and standard setting.