Journal of Experimental Psychology: Applied, 2013, Vol. 19, No. 4, 383– 402

A Little Bias Goes a Long Way: The Effects of Feedback on the Strategic Regulation of Accuracy on Formula-Scored Tests

Michelle M. Arnold, Flinders University

Philip A. Higham, University of Southampton

Beatriz Martín-Luengo, University of the Basque Country (UPV-EHU)

Michelle M. Arnold, School of Psychology, Flinders University, Ade- laide, Australia; Philip A. Higham, Department of Psychology, University of Southampton, Southampton, UK; Beatriz Martín-Luengo, Faculty of Psychology, University of the Basque Country (UPV-EHU), San Se- bastián, Spain.

We thank Karlos Luna for programming assistance. Preparation of this article was supported by a grant from the British Academy. Portions of this research were presented at the 47th Annual Meeting of the Psychonomic Society, Houston, TX.

Correspondence concerning this article should be addressed to Michelle M. Arnold, School of Psychology, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia. E-mail: michelle.arnold@flinders.edu.au

Under formula-scoring rules for multiple-choice exams, a penalty is applied to incorrect responses to reduce noise in the observed score. To avoid the penalty individuals are allowed to “pass,” and therefore they must be able to strategically regulate the accuracy of their reporting by deciding which and how many questions to answer. To investigate the effect of bias within this framework, Higham (2007) introduced bias profiles, which show the score obtained under formula scoring (corrected score) as a function of the omission rate. Bias profiles estimate the optimal number of questions that should be answered to maximize the corrected score (i.e., *optimal bias*). Our initial research showed that individuals tend to be too conservative when setting reporting criteria, “omitting” too many answers. The present three experiments introduced a feedback manipulation whereby participants were informed of the optimal omission rate after completing a test and asked to alter their reporting decisions accordingly. This feedback and concomitant alteration of reporting decisions led to improved corrected scores on true/false (Experiment 1), 2-alternative tests (Experiments 2), and 4-alternative tests (Experiment 3). Importantly, corrected scores at optimal bias also were higher than at forced-report for both true/false and 2-alternative tests. Furthermore, in Experiment 3, feedback based on one test improved scores on a second test, and participants were more likely to perform optimally on a third test without feedback. These effects suggest that optimal-bias feedback may have long-term effects and generalize to new tests.

Keywords: bias profiles, formula scoring, metacognition, strategic regulation of accuracy, type-2 signal detection theory

A great deal of research in metacognition has focused on people’s ability to predict their performance on future tests. For example, researchers have examined the extent to which judgments about how well paired associates have been learned predict subsequent performance on an actual memory test (e.g., Dunlosky & Hertzog, 2000). Much less attention has been paid specifically to the role metacognition plays in determining performance on a current test. Recently, however, Higham and colleagues (e.g., Higham, 2007; Higham & Arnold, 2007a, 2007b) have examined metacognition in multiple-choice testing scenarios in which students are given the opportunity to pass (i.e., leave questions unanswered). The motivation to pass on these tests typically is invoked by *formula scoring*, which involves a point system. Specifically, in most instantiations of formula scoring, correct responses earn points, but incorrect responses incur a penalty (e.g., Muijtjens, van Mameren, Hoogenboom, Evers, & van der Vleuten, 1999; Thurstone, 1919). However, the penalty that is applied to incorrect responses can be avoided by omitting answers; examinees can choose to omit responses to questions if they are unsure of the correct answer, which usually neither gains nor loses points. To maximize their corrected scores, students must strategically regulate both which and how many responses to report. Such regulation must find a balance between reporting too much and receiving penalties for low-quality answers (sunk cost), and reporting too little and missing out on points for high-quality answers (opportunity cost).

Formula scoring is an alternative to the traditional *number-right* scoring system in which test-takers are instructed to answer all items and the test is scored by assigning one point for every correct response (i.e., no penalty for errors). Because formula scoring has a report/omit option, it introduces the strategic regulation of accuracy, which is not present in the number-right system. Therefore, it is important to consider issues surrounding metacognitive monitoring and control when investigating performance on formula- scored tests. Currently, there are two frameworks that are used to investigate the strategic regulation of accuracy, the first of which is Koriat and Goldsmith's (1996; Goldsmith & Koriat, 2008 monitoring-control framework. The second is a type-2 signal detection theory (SDT) framework that has been applied to a testing context by Higham (2007; see also Higham & Arnold, 2007a, 2007b; Lueddeke & Higham, 2011), and which is the framework adopted in this study (see Higham, 2011, and Goldsmith, 2011, for a comparison and discussion of the two approaches).

To investigate accuracy regulation in test-taking situations un- der both frameworks, it is necessary to know the answers that students omitted. One method for obtaining these answers is sim- ply to have students return to questions initially left unanswered and ask them to provide best guesses (e.g., Bliss, 1980; Cross & Frary, 1977; Ebel, 1968; Higham, 2007, Experiment 1; Muijtjens et al., 1999; Sax & Collet, 1968; Sherriffs & Boomer, 1954; Slakter, 1968a, 1968b). However, this two-pass procedure is not ideal because students have two attempts at answering questions for which responses were initially omitted (but only one attempt at other questions on the test), which introduces a systematic processing-time difference between the question sets. To avoid this criticism, Higham (2007, Experiment 2) developed a one-pass procedure that required both a response to every question and an assignment of each response to either a "go for points" category (where correct and incorrect responses garnered points and penal- ties, respectively) or a "guess" category (where both correct and incorrect responses earned 0 points). The "go for points" and "guess" category assignments were considered analogous to re- porting and omitting responses, respectively. Higham found that the one- and two-pass procedures produced very similar results, but the one-pass procedure avoids the criticism that not all questions on the test receive equal consideration. For this reason, Higham's one-pass procedure is adopted in the current experiments, and for ease of exposition we will use the terms "report" and "omit" in many places throughout the rest of the study to refer to "Go for Points" and "Guess" category assignments, respectively (even though, technically, answers are never omitted with the one-pass procedure).

Beyond avoiding the potential criticisms of the two-pass procedure, the one-pass method provides a means to gauge metacognitive monitoring (resolution) and bias using type-2 SDT. Because responses are assigned to the “guess” category rather than omitted, they can be scored as correct or incorrect. Knowing the correctness of these responses means that the type-2 hit rate (HR) and false alarm rate (FAR) can be calculated. As shown in Table 1, the HR is equal to the number of correct answers placed in the “go for points” column divided by the total number of correct responses (*a*/[*a* + *c*]), whereas the FAR is equal to the number of incorrect responses in the “go for points” column divided by the total number of incorrect responses (*b*/[*b* + *d*]). Once the HR and FAR are known, a discrimination index (e.g., *d*=) can be calculated. As detailed below, a key feature of type-2 SDT is that the discrimination index is a measure of metacognitive resolution.

Although resolution typically has been measured with the Goodman-Kruskal gamma coefficient (Goodman & Kruskal, 1954)—an ordinal measure of correlation recommended by Nelson (1984)—it recently has been shown to have a number of poor qualities and SDT indices have been suggested as an alternative (e.g., Higham, 2007, 2011; Luna, Higham, & Martin-Luengo, 2011; Masson & Rotello, 2009; Rotello, Masson, & Verde, 2008). It is also important to distinguish resolution from *calibration*, another type of metacognitive accuracy. Whereas resolution (or *relative metacognitive accuracy*) is a measure of the degree to which people can discriminate the correctness of their own responses, *calibration* (or *absolute metacognitive accuracy*) is the degree to which the units of measurement on a metacognitive scale match actual performance. Throughout this paper, we only will be examining resolution, because it is the measure that corresponds to discrimination from SDT

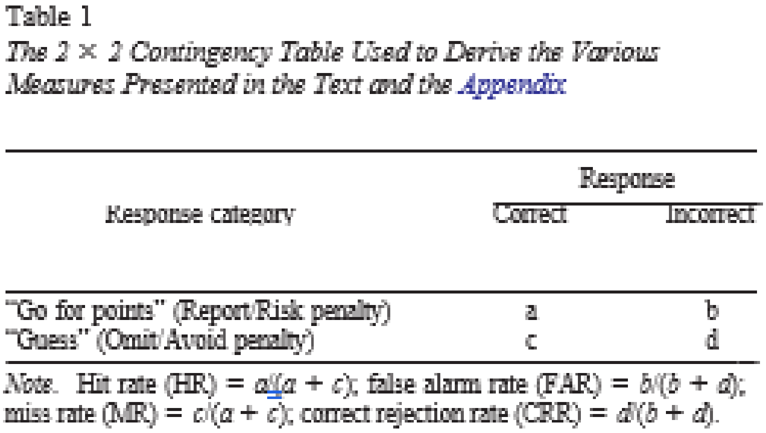

Table 1

The 2 × 2 Contingency Table Used to Derive the Various Measures Presented in the Text and the Appendix

Response category	Response	
	Correct	Incorrect
"Go for points" (Report/Risk penalty)	a	b
"Guess" (Omit/Avoid penalty)	c	d

Note. Hit rate (HR) = *a*/(*a* + *c*); false alarm rate (FAR) = *b*/(*b* + *d*); miss rate (MR) = *c*/(*a* + *c*); correct rejection rate (CRR) = *d*/(*b* + *d*).

Type I Versus Type II SDT

To understand better how the type-2 SDT framework provides a measure of resolution, refer to Figure 1. For all SDT tasks, the goal of the observer is to discriminate between trials that contain only noise (N) and trials that contain a signal-plus-noise (S; Green & Swets, 1966; Macmillan & Creelman, 2005). To accomplish this task, the observer must adopt a criterion along a dimension of sensed intensity of the signal, however that dimension might be defined. In the type-1 SDT case depicted in Figure 1A, this dimension is the sensed intensity of a stimulus dimension such as “familiarity” if the task is old/new recognition, or “brightness” if it is a perceptual task. If the sensed intensity of a given trial is at or above the value of the criterion, the observer will respond “yes, the signal is present,” otherwise “no, the signal is not present.” For example, in an old/new recognition task, “yes” and “no” responses would translate into “old” and “new” responses, respectively.

In the model depicted in Figure 1A, the measure of discrimination—shown in the figure as *d*=—is an index of the observer’s ability to discriminate between different types of stimulus trials, whereas the criterion is a measure of an individual’s tendency to respond “yes.” An individual who frequently responds “yes” would have a criterion to the left of the intersection point of the two distributions and is said to have a liberal bias. Conversely, someone who is less willing to responding “yes” would have a criterion shifted to the right of the intersection point (as in Figure 1A) and is said to have conservative bias.

In a type-1 discrimination task, S and N trials typically are experimenter-determined (e.g., the experimenter decides which test items are lures vs. targets in recognition memory experiments). However, in a type-2 discrimination task, it is the observers themselves who define the S and N trials; in particular, observers must discriminate between their own correct (S) and incorrect (N) responses. This critical change can alter the nature of the S and N distributions and the underlying dimension over which the discrimination is made (Figure 1B). That is, the S and N distributions are correct and incorrect candidate responses, respectively, and the dimension is the sensed intensity of correctness. Importantly, these alterations mean that observers must decide whether there is enough “sensed correctness” to report an answer and the discrimination index (e.g., *d*=) in a type-2 task becomes a metacognitive index of resolution (akin to a confidence-accuracy correlation), rather than a bias-free measure of accuracy, as with type-1 SDT. Finally, the type-2 criterion is the observer’s tendency to report an answer. For a more detailed discussion of SDT and how it relates to formula scoring, see Higham (2007) and Higham and Arnold (2007a, 2007b).

Corrected Versus Raw Scores

Formula scoring originally was introduced to remove error (guessing) variance from the observed raw score to yield a purer measure of knowledge/aptitude, and thus enhancing the reliability and validity of the test. However, the SDT framework seriously questions this basic classical-testing tenet. From the SDT perspective, the corrected score is a highly complex measure of performance that is influenced not just by knowledge/aptitude but three separable parameters: (1) knowledge/aptitude - the uncorrected proportion of items on the test that are correct after all questions have been answered (*f*), (2) resolution - participants’ ability to monitor the correctness of their own candidate answers (e.g., *d*=), and (3) bias - the tendency to report/assign answers to the “go for points” versus omit/assign them to the “guess” category. The parameter *f* is not the raw score (or “number right”) per se, but the raw score divided by the number of items on the test. For example, an examinee who answers 40 of 50 questions correctly under the number-right scoring system would have a raw score of 40, whereas *f* = 40/50 = .80. Thus, *f* contains the same information as the raw score about knowledge/aptitude, but it has the advantage that it can be directly compared between tests of differing length. Further, it is important to note that these three separate parameters are respectively analogous to the retrieval, monitoring, and control parameters in Koriat and Goldsmith’s (1996) framework.

From the SDT perspective, any observed difference between the raw score on the one hand and the corrected score on the other is attributable to *imperfect resolution* and/or *imperfect criterion set- ting*. Students who monitor their knowledge perfectly and who set a criterion at an optimal level will report all of their correct answers (thus avoiding opportunity costs) and omit all of their incorrect ones (thus avoiding sunk costs), producing identical raw and corrected scores. However, even students with perfect monitoring who do not set an optimal criterion, or students who set a criterion optimally but have imperfect monitoring, will produce a corrected score that is lower than the raw score. Thus, to the extent that any difference between the scores is observed, it highlights room for improvement on one, the other, or both of these performance parameters

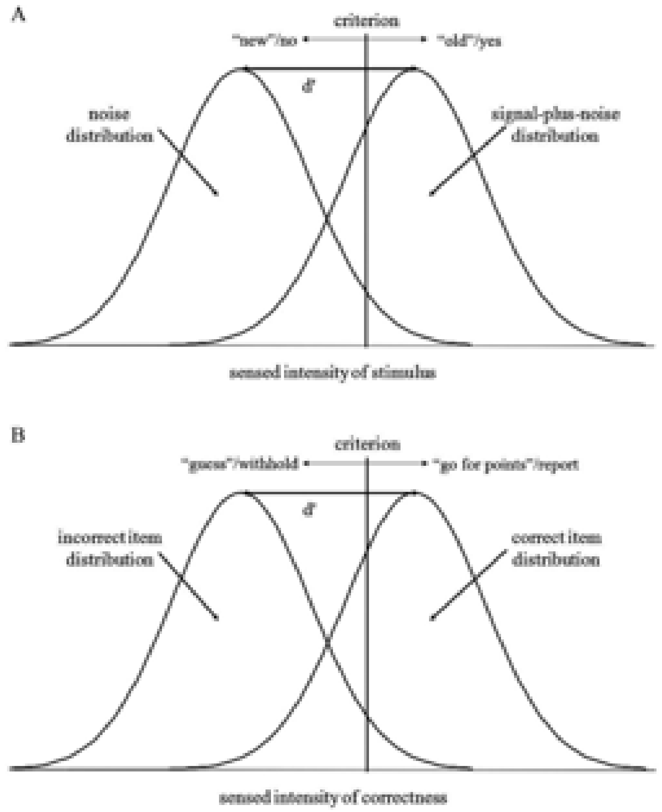
A
criterion
"new"/no
"old"/yes
d'
noise distribution
signal-plus-noise distribution
sensed intensity of stimulus
B
criterion
"guess"/withhold
"go for points"/report
d'
incorrect item distribution
correct item distribution
sensed intensity of correctness

Figure 1. Type-1 (A) and type-2 (B) signal-detection models with equal-variance, normal distributions.

Closing the Gap Between Raw and Corrected Scores: Bias Profiles

In the type-1 SDT domain, much of the research has focused on how participants adjust their response criterion to maximize performance (e.g., Bisseret, 1981; Cañal-Bruland & Schmidt, 2009; Maddox, 2002; Meissner & Kassin, 2002; Parasuraman, 1985). For example, in recognition memory research, when targets are strengthened by repeating them during study, participants tend to adopt a more conservative old/new criterion (e.g., Bruno, Higham & Perfect, 2009; Stretch & Wixted, 1998). Doing so maximizes accuracy on the test as a whole, compared with a situation in which the criterion remains static after targets have been strengthened.

The current article builds on this previous work examining optimal criterion setting in two ways. First, we examined optimal placement of the type-2 report/omit criterion rather than the type-1 yes/no criterion (see also Goldsmith & Koriat, 2008; Lueddeke & Higham, 2011). This distinction may be important because the cues that people rely on to strategically control their criterion placement may be different when making response-contingent (type-2) decisions rather than stimulus-contingent (type-1) ones. Second, to fully examine how the corrected score varies with bias for particular students, we used *bias profiles*, which were introduced by Higham (2007; see also Higham & Arnold, 2007a, 2007b; Lueddeke & Higham, 2011). A bias profile is a plot of the corrected score as a function of the number of omitted responses. A critical feature of a bias profile is that it produces a measure of *optimal bias*, that is, the number of items students should omit to achieve their maximum-corrected score. The mathematical details and calculations for how to generate bias profiles are shown in the Appendix, but for more detailed discussion we refer readers to Higham (2007); Higham and Arnold (2007a, 2007b), and Lued- deke and Higham (2011). In short, bias profiles are sensitive to three factors: (1) resolution (*d*=), (2) the raw, uncorrected proportion of items correctly answered (*f*), and (3) the penalty for incorrect responses (*p*). These parameters can then be fixed in the equations and the FAR varied between 0 and 1 to produce values of the corrected score, which are then plotted against the corresponding omission rate.

Bias profiles are a straightforward extension of some basic SDT principles and are analogous to the Receiver Operating Characteristic (ROC) curves that are unique to SDT. Some examples of bias profiles for a two-alternative-forced-choice (2AFC) test with fixed parameters *d*= (1.25) and *p* (1.00), and three levels of *f* (.50, .70, and .85) are presented in Figure 2. Unsurprisingly, note that as *f* increases, so does the corrected score. More interesting, however, is the observation that as *f* increases, the proportion of omissions needed to achieve the maximum possible corrected score de- creases. For example, if *f* equals .50, examinees should omit 50% of their responses to achieve a maximum-corrected score of .23 (bottom curve). Conversely, if *f* increases to .85, examinees need only omit 5% of their answers to attain the maximum-corrected score of .72 (top curve). The *d*= and *p* parameters also affect the shape of bias profiles and optimal bias, sometimes in ways that are somewhat counterintuitive (see Higham & Arnold, 2007b, for specific examples). Nonetheless, bias profiles provide a means to determine, for a particular student on a particular test, exactly how many questions should be omitted so that the maximum-corrected score can be achieved.

If students are setting criteria optimally on an exam, then their *actual bias* (i.e., the proportion of answers they actually omit) should match their optimal bias, as calculated by the bias profile. Higham and Arnold (2007a) tested this relationship between optimal and actual bias using classroom data; bias profiles were created for each student who wrote three formula-scored exams (i.e., under "go for points"/"guess" instructions) in an introductory psychology course. The results demonstrated that, across all three exams, students did not perform at their optimal bias. Specifically, the overall mean optimal-bias score showed that students should have omitted 10% of the questions, whereas their mean actual bias revealed that they were too conservative and omitted 25% of their responses—a phenomenon we refer to as *underconfidence*. The term "underconfidence" is also used to describe cases in which rated confidence underestimates actual performance in calibration research. For example, participants who rate a group of items as 60% likely to be recalled but who later recall 80% of them are described as underconfident. However, throughout this study, we use the term to refer to *ultraconservatism*, that is, omitting too many answers such that the corrected score suffers on formula- scored tests. More importantly, the discrepancy Higham and Arnold found between optimal and actual bias remained almost constant across the three exams, in that having experience with writing formula-scored exams (and receiving feedback in the form of corrected scores) did not teach students how to optimize their test-taking strategy and reduce their omission rate by the final test. Although Higham and Arnold's (2007a) data support the idea that guessing in the formula-scoring system is not random and that students underestimate their knowledge (i.e., by failing to report high-quality responses), there were some methodological issues that may hamper a clear interpretation of the results. For example, on all three exams a small bonus (.25) was given for any correct responses that were omitted and therefore critics may argue that this bonus was an incentive to omit responses, which resulted in an inflated measure of the actual bias. Similar results have been found when no bonus was offered for incorrect omissions (Higham, 2007), but it is important to note that, in general, a bonus for correct omissions is a deviation from the typical formula-scoring system. Another potential issue that may cloud generalizability from the data is that the penalty was higher on Test 1 (.50) than both Test 2 and 3 (.33); starting students with a more severe penalty (i.e., one that is higher than the typical guessing correction) may have pushed them to be more conservative than they other- wise would have been, although it is not clear why they would have remained too conservative across all three tests. Regardless, these two issues, along with the fact that the design was correlational, are important considerations when trying to reach concrete conclusions from the results. Therefore, the goal of the present research both was to replicate and to build on the Higham and Arnold findings in an experimental context.

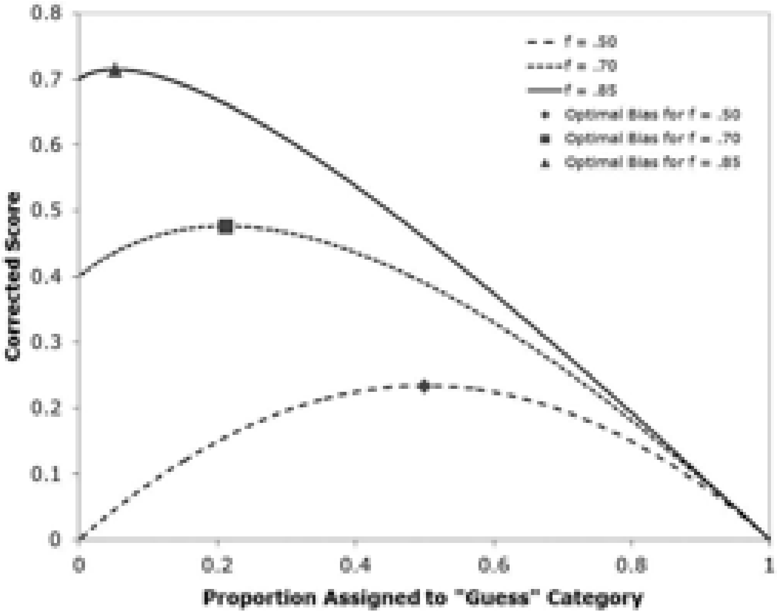

Figure 2. Hypothetical bias profiles calculated for a fixed *d*= and p of 1.25 and .33, respectively, and three different levels of f (.50, .70, and .85).

Overview of the Experiments

In three experiments, we administered general-knowledge tests and participants completed each test under the one-pass procedure described above (i.e., answer all questions but assign each to either a "go for points" or "guess" category). The first overall goal of these experiments was to replicate in a controlled, experimental context Higham and Arnold's (2007a) critical results—namely, that participants are underconfident in their knowledge and there- fore omit too often. The second aim of the present set of experiments was to extend the previous research by adding an optimal- bias feedback variable. Although students in Higham and Arnold's (2007a) study received feedback in the form of test results (i.e., they were given their Test 1 and 2 exams to review, along with the correct answers to the questions), they received no advice regarding test-taking strategy or optimal bias. In contrast, in Experiment 1 and 2 of the current research, participants received feedback at the end of each test and subsequently were asked to adjust their criterion setting accordingly. Specifically, bias profiles were calculated at the end of each test for the participants and they were then informed of their optimal bias (i.e., the proportion of answers that should be omitted to achieve the maximum-corrected score); participants were then required to go back over the test and redistribute their answers between the "go-for-points" (report) and "guess" (omit) categories until optimal bias was achieved. For example, suppose the optimal bias computation on a student's completed test indicated that 10% of the questions should have been omitted to reach the maximum-corrected score, but 20% were actually omitted. This participant would be informed of his or her actual bias and required to adjust the "guess"/"go for points" category assignments until the omission rate was reduced to 10%. Importantly, the answers themselves could not be changed, only the decision on whether to "guess" or "go for points." Thus, if there was any increase in the corrected score that resulted from the redistribution of answers, we could be sure that it was specifically optimal criterion setting that was resulting in the corrected-score increase and not answer changing.

Experiment 3 was similar in nature to the first two experiments, but there were two experimental groups, *control* and *educated*. Participants in the control group completed three tests under formula-scoring instructions, and they received no feedback on their performance at any point during the experiment. The partic- ipants in the educated group completed the same three tests, but after finishing Test 2 they were informed of their optimal-bias score from *Test 1* and told to return to their *second* test and adjust assignment of items to the "go for points" and "guess" categories accordingly. Participants then were required to write Test 3 with- out further guidance. This methodology was used to explore two important questions. First, can the optimal bias measure from one test be applied to a subsequent test in order to improve perfor- mance (i.e., significantly increase the corrected scores on the subsequent test compared to initial performance)? Second, is there a lasting benefit of feedback in terms of criterion setting on a third test? That is, after feedback (based on Test 1) has indicated the appropriate number of answers to omit on Test 2, are students less underconfident on a final test?

The third aim of the current experiments was to investigate the efficacy of normative advice given to students that is derived from judgment and decision theory. Budescu and Bar-Hillel (1993) noted that, for a rational test taker, reporting answers should always equal or increase the corrected score compared to omitting; that is, there should never be a score advantage for a test with omissions compared with one without. Indeed, as noted above, a common finding in the testing literature is that providing "best guesses" to questions for which answers were originally omitted tends to improve the test score (e.g., Bliss, 1980; Cross & Frary, 1977; Ebel, 1968; Higham, 2007, Experiment 1; Muijtjens et al., 1999; Sax & Collet, 1968; Sherriffs & Boomer, 1954; Slakter, 1968a, 1968b). Thus, the suggestion seems to be that students should ignore warnings to omit answers if they are unsure and simply treat formula-scored tests as if they were number-right tests and never leave a question blank. On the other hand, although bias profiles sometimes predict that a zero omission rate will maximize the corrected score, often they have an inverted-U shape such that optimal bias is greater than zero (see Figure 2). Thus, judgment and decision theory on the one hand and SDT on the other appear to sometimes give conflicting advice as to how students should behave when writing formula-scored tests.

To investigate this important matter, we compared corrected scores given at SDT-based optimal bias with corrected scores at forced report. If the predictions for the rational test taker from judgment and decision theory are correct, then the forced-report score should be equal to, or exceed, the score at optimal bias. Conversely, if the inverted U-shaped predictions of SDT are cor- rect, then the corrected score at optimal bias should exceed that at forced report (assuming optimal bias is not equal to zero).

Experiment 1

Method

Participants. Twenty-four University of Southampton under- graduates participated in exchange for £5.

Design and materials. Three general knowledge tests were constructed, each of which contained 26 true/false (T/F) questions gathered from a variety of sources (e.g., Nelson & Narens, 1980). Three sets of 26 questions (question set: *A*, *B*, and *C*) were constructed to counterbalance the items across the tests, which resulted in six between-subjects counterbalancing conditions. The questions sets were counterbalanced so that each set appeared equally often across participants in Test 1, 2, or 3. Testing was computerized and the response screen for each test question in- cluded the following (1) The T/F question (e.g., "Omega is the last letter of the Greek alphabet"), (2) separate "Go for points" and "Guess" columns containing the "True" and "False" response options and corresponding points system (i.e., "+1/-1" for "Go for points" and "0 points" for "Guess") and (3) a section to record confidence in the chosen response on a scale from 1 to 6 (with 1 being least confident and 6 being most confident). For responses assigned to the "Go for points" category, the typical correction factor (penalty for errors) was applied (i.e., 1/[number of alterna- tives — 1]), which for a T/F test with two alternatives, was equal to 1.

Procedure. All participants were tested in individual work-

stations. To ensure that participants understood the structure of the tests, they were given a set of instructions to read before starting Test 1, which were verbally reiterated by the experimenter. Both the written instructions and the experimenter informed participants that they had two options when answering a question that would indicate whether or not they wanted their response to count in the point tally. More specifically, they were informed that if they believed that they knew the correct answer to a question then they should select the answer (i.e., click either "True" or "False") in the "Go for points" column—if they selected the correct answer they would receive 1 point, but if they gave an incorrect response they would be penalized 1 point. Alternatively, if they believed that they did not know the correct answer to a question (i.e., that they were guessing) then they were told to select their True/False answer in the "Guess" column, in which case no points would be gained or lost. Finally, participants were told that, regardless of whether they chose to "go for points" or "guess," they had to rate their confidence in their chosen response using the scale of 1 to 6. At the end of each test the experimenter calculated participants' optimal-bias scores (i.e., the number of responses they should have omitted to achieve their maximum-corrected score). All participants were then given optimal-bias feedback. That is, participants were told that their optimal bias had been calculated for the test, and that they were restricted to this number. If this value differed from actual bias, they were required to redistribute their answers between the "Guess" and "Go for points" columns. Specifically, if there were too many items in the "Guess" column (i.e., optimal bias < actual bias), then they should choose answers that were likely to be correct from that category to move to the "Go for points" category. However, if there were too many items in the "Go for points" category (i.e., optimal bias > actual bias) then they should choose answers that were likely to be incorrect from that category to move to the "Guess" category. Additionally, partici- pants were told that they were not permitted to make any changes to their actual responses or their confidence ratings. Once partic- ipants had the required amount of responses placed in the "Guess" and "Go for points" columns (i.e., had achieved their optimal bias) they were allowed to move on to the next test.

Results and Discussion

Corrected-score differences. There are four different types of corrected scores that are of interest: (1) the highest corrected score that was theoretically possible, as determined by bias profiles (*maximum-corrected score*), (2) the observed corrected score for a formula-scored exam (*corrected score*), (3) the observed corrected score after participants were instructed to redistribute their answers between the "go for points" and "guess" categories to reach optimal bias (*corrected-feedback score*), and (4) the observed corrected score that would be achieved if participants were in- structed to report all answers (*corrected-forced score*). Overall, if the current data replicate the previous findings (Higham & Arnold, 2007a), then the mean corrected score will be significantly lower than the maximum-corrected score. Further, if the feedback ma- nipulation was successful then the corrected-feedback score should be higher than the corrected score. However, even if optimal-bias feedback produces a higher corrected score, a critic may argue that this increase came about because students simply answered more questions and that the better/easier strategy to teach participants is that they should never omit responses; that is, it is possible that having participants report all responses would have been just as effective (or more effective) than having them perform at optimal bias (Budescu & Bar-Hillel, 1993). Therefore, if there is a signif- icant increase in corrected scores attributable to optimal-bias feed- back, it is important to demonstrate that this increase is higher than if the participants had been given the easier strategy of reporting all answers (i.e., corrected-feedback > corrected-forced).

The four different corrected scores across the three tests are presented in Table 2. A 4 (*score*: maximum-corrected, corrected, corrected-feedback, corrected-forced) × 3 (*test*: Test 1, Test 2, Test 3) repeated-measures ANOVA demonstrated a main effect of score, *F*(3, 69) = 19.19, *MSE* = .003, h^2 = .46, *p* < .001,$_p$but no

main effect of test and no interaction between test and score, *F*s ≤ 1.06, *p*s ≥ .39. Planned follow-up comparisons showed that the maximum-corrected score (*M* = .50, *SEM* = .02) was significantly higher than the corrected score (*M* = .43, *SEM* = .02), *t*(23) = 8.36, *p* < .001, the corrected-feedback score (*M* = .48, *SEM* =

.03), *t*(23) = 4.27, *p* < .001, and the corrected-forced score (*M* =

.45, *SEM* = .03), *t*(23) = 5.80, *p* < .001. Importantly, the corrected-feedback score was significantly greater than both the corrected score, *t*(23) = 4.54, *p* < .001, and the corrected-forced score, *t*(23) = 2.52, *p* = .02. There was no difference between the corrected score and the corrected-forced score, *t*(23) = 1.38, *p* =

.18.

Resolution. The proportions of correct versus incorrect re- sponses that were reported correspond to the type-2 SDT HR versus FAR, respectively, which can be used to index relative monitoring accuracy or resolution. Specifically, if the difference between these rates is high, participants are using the report option to successfully monitor the accuracy of their own responses. Con- versely, if the difference is low, their resolution is poor. To investigate resolution, a 2 (*feedback*: before, after) × 2 (*response*: HR, FAR) × 3 (*test*: Test 1, Test 2, Test 3) within-subjects ANOVA was performed on the proportion of responses placed in the "go for points" column (see Table 3 for the means). There was a main effect both for feedback, *F*(1, 23) = 72.74, *MSE* = .05, h^2 = .76, *p* < .001, and response, *F*(1, 18) = 56.82, *MSE* = .07,

p

h^2 = .71, *p* < .001. The main effect of feedback arose because

participants were underconfident prefeedback (i.e., too conserva- tive); the HRs and FARs were lower before (*M* = .61, *SEM* = .03) than after (*M* = .83, *SEM* = .02) the feedback was provided across all three tests. As anticipated, the main effect of response occurred because the HR (*M* = .84, *SEM* = .01) was higher than the FAR rate (*M* = .60, *SEM* = .03), showing that resolution was reason- ably good. Of more interest, though, was an interaction between feedback and response, *F*(1, 23) = 33.28, *MSE* = .01, h^2 = .59, *p* < .001, which occurred because *p* the difference between pre- and postfeedback for the HRs versus the FARs showed that feedback led to a larger increase in the FAR (*M*diff = .28, *SEM* = .03) than the HR (*M*diff = .16, *SEM* = .02). No other main effect or interactions were significant, *F*s ≤ 1.93, *p*s ≥ .16.

Optimal versus actual bias. Bias profiles were created for each participant for the three tests, and optimal and actual bias are reported in Table 4. Optimal and actual bias reflect the proportions of items that should have been omitted (according to the bias profiles), and that actually were omitted (before feedback), respec- tively. These data were analyzed in a 2 (*bias type*: optimal, actual) × 3 (*test*: Test 1, Test 2, Test 3) within-subjects ANOVA. The analysis revealed a main effect of bias type, *F*(1, 23) = 82.90, *MSE* = .02, h ^{2} = .78, *p* < .001, which arose because participants were too conservative in their responding; that is, actual bias (*M* =.32, *SEM* = .02) was significantly higher than optimal bias (*M* =.12, *SEM* = .02). There was no main effect of test and no interaction between test and bias type, *F*s ≤ 1.34, *p*s ≥ .26.

Confidence ratings. The mean confidence ratings across the three tests are shown in Table 5. A 2 (*response*: correct, incorrect) × 2 (*report decision*: go for points, guess) × 3 (*test*: Test 1, Test 2, Test 3) within-subjects ANOVA was conducted on the mean ratings. The results showed main effects of response, *F*(1, 15) = 38.53, *MSE* = .29, h^2 = .72, *p* < .001, and report decision, *F*(1, 15) = 122.45, *MSE* = 1.90, h^2 = .89, *p* < .001. The main effects of response and report decision occurred because, as expected, participants assigned higher confidence ratings to correct (*M* = 3.65, *SEM* = .17) than incorrect (*M* = 3.17, *SEM* = .20) items, and to items placed in the "go for points" (*M* = 4.51, *SEM* = .20) than "guess" (*M* = 2.30, *SEM* = .21) column. There also was an interaction between response and report decision, *F*(1, 15) = 26.18, *MSE* = .32, h ^{2} = .64, *p* < .001; the difference in confidence ratings between correct and incorrect responses was larger for items placed in the "go for points" column (M$_{diff}$ = .90, *SEM* = .14) than for items placed in the “guess” column (*M*diff = .07, *SEM* = .07). There was no main effect of test, and no other interactions were significant, all *F*s ≤ 1.10, *p*s ≥ .35.

Summary. The results from Experiment 1 replicated Higham and Arnold’s (2007a) key findings that individuals tend to be underconfident in their level of knowledge when answering formula-scored tests, and that overall they adopt a criterion that is too conservative (e.g., maximum-corrected score > corrected score; see also Budescu & Bar-Hillel, 1993; Koriat & Goldsmith, 1996). These finding were obtained despite the change in the nature of the test from 4AFC used in Higham and Arnold’s research to the T/F format used here. Further, the confidence data demonstrated that self-reported confidence was a predictor of accuracy for the reported items: Correct reported answers were given a significantly higher confidence rating than incorrect re- ported answers.

The present study also showed that providing feedback about optimal bias and instructing participants to distribute their answers between the “Go for points” and “Guess” categories accordingly led to a significant improvement in their test scores. Interestingly, the resolution data demonstrated that the optimal-bias feedback led to a higher increase in the FAR than the HR, which may suggest one reason why examinees tend to be too conservative; specifically, they may have some intuition that responding at optimal bias rather than at their prefeedback bias had forced them to report a considerably larger proportion of their errors (i.e., they are particularly sensitive to their FAR). Despite the large effect of feedback on the FAR, participants moved more correct answers (*M* = 2.99, *SEM* = .37) than incorrect answers (*M* = 1.85, *SEM* = .27) to the “go for points” category in response to optimal-bias feedback, which explains why their corrected score increased.

Although participants moved more correct answer to the “go for points” category in response to optimal-bias feedback, the careful reader may wonder how participants’ corrected score could in- crease after redistributing their responses in response to optimal- bias feedback if the redistribution caused a greater increase to the FAR than the HR; should that not mean that the penalty would overshadow any points gained, leading to a reduction of the corrected score rather than an increase? The reason that there is an increase in the corrected score highlights again how the corrected score is a complex measure made up of several components; in particular, because *f* is above .50, a change to the *b* frequency in Table 1 (false alarms) will have a greater impact on the FAR (*b*/[*b* + *d*]) than a comparable or even larger change to the *a* frequency (hits) will have on the HR (*a*/[*a* + *c*]). In other words, because there are fewer incorrect (*b* + *d*) than correct (*a* + *c*) responses, the FAR is more susceptible to change than the HR.

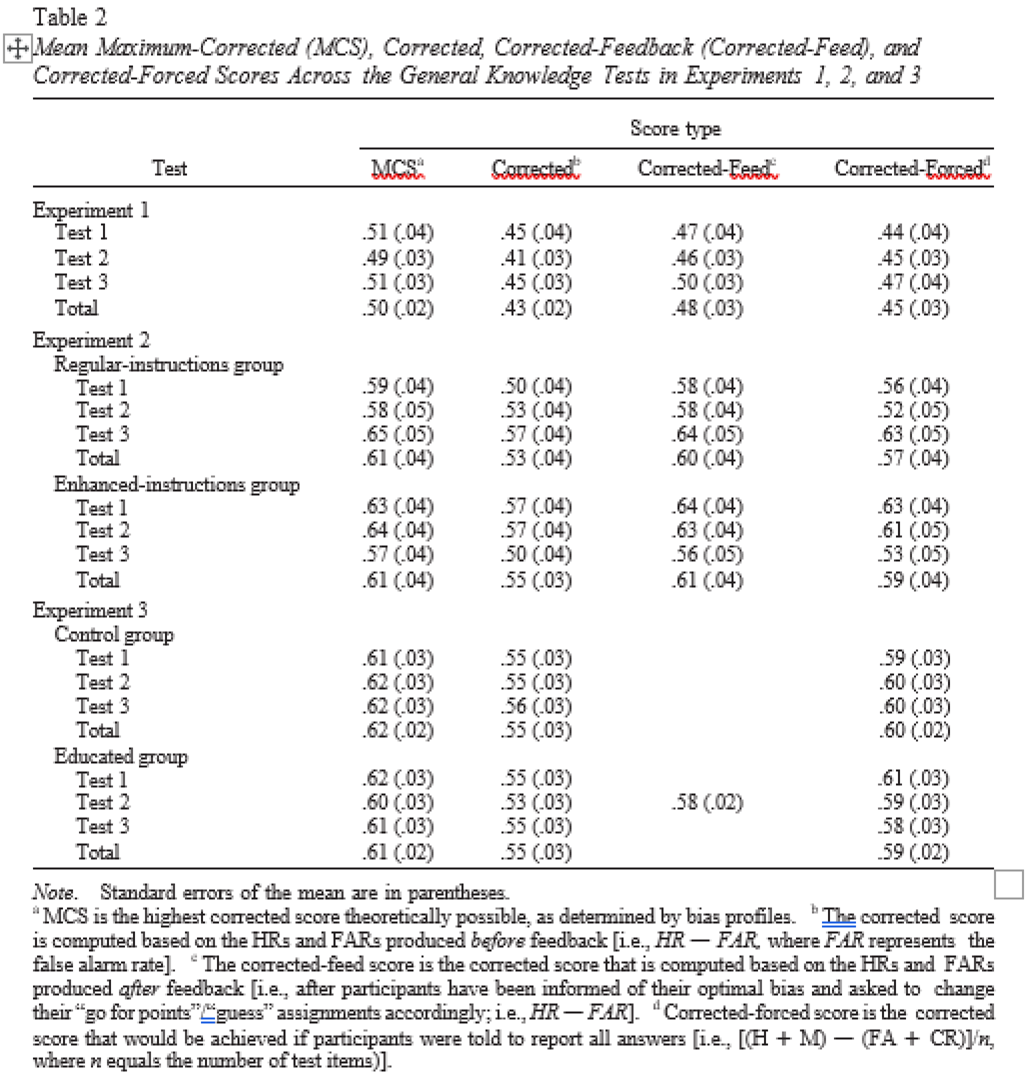

Table 2

Mean Maximum-Corrected (MCS), Corrected, Corrected-Feedback (Corrected-Feed), and Corrected-Forced Scores Across the General Knowledge Tests in Experiments 1, 2, and 3

Test	Score type			
	MCS[a]	Corrected[b]	Corrected-Feed[c]	Corrected-Forced[d]
Experiment 1				
Test 1	.51 (.04)	.45 (.04)	.47 (.04)	.44 (.04)
Test 2	.49 (.03)	.41 (.03)	.46 (.03)	.45 (.03)
Test 3	.51 (.03)	.45 (.03)	.50 (.03)	.47 (.04)
Total	.50 (.02)	.43 (.02)	.48 (.03)	.45 (.03)
Experiment 2				
Regular-instructions group				
Test 1	.59 (.04)	.50 (.04)	.58 (.04)	.56 (.04)
Test 2	.58 (.05)	.53 (.04)	.58 (.04)	.52 (.05)
Test 3	.65 (.05)	.57 (.04)	.64 (.05)	.63 (.05)
Total	.61 (.04)	.53 (.04)	.60 (.04)	.57 (.04)
Enhanced-instructions group				
Test 1	.63 (.04)	.57 (.04)	.64 (.04)	.63 (.04)
Test 2	.64 (.04)	.57 (.04)	.63 (.04)	.61 (.05)
Test 3	.57 (.04)	.50 (.04)	.56 (.05)	.53 (.05)
Total	.61 (.04)	.55 (.03)	.61 (.04)	.59 (.04)
Experiment 3				
Control group				
Test 1	.61 (.03)	.55 (.03)		.59 (.03)
Test 2	.62 (.03)	.55 (.03)		.60 (.03)
Test 3	.62 (.03)	.56 (.03)		.60 (.03)
Total	.62 (.02)	.55 (.03)		.60 (.02)
Educated group				
Test 1	.62 (.03)	.55 (.03)		.61 (.03)
Test 2	.60 (.03)	.53 (.03)	.58 (.02)	.59 (.03)
Test 3	.61 (.03)	.55 (.03)		.58 (.03)
Total	.61 (.02)	.55 (.03)		.59 (.02)

Note. Standard errors of the mean are in parentheses.
[a] MCS is the highest corrected score theoretically possible, as determined by bias profiles. [b] The corrected score is computed based on the HRs and FARs produced *before* feedback [i.e., *HR* — *FAR*, where *FAR* represents the false alarm rate]. [c] The corrected-feed score is the corrected score that is computed based on the HRs and FARs produced *after* feedback [i.e., after participants have been informed of their optimal bias and asked to change their "go for points"/"guess" assignments accordingly; i.e., *HR* — *FAR*]. [d] Corrected-forced score is the corrected score that would be achieved if participants were told to report all answers [i.e., [(H + M) — (FA + CR)]/*n*, where *n* equals the number of test items)].

underconfident in their level of knowledge when answering formula-scored tests, and that overall they adopt a criterion that is too conservative (e.g., maximum-corrected score > corrected score; see also Budescu & Bar-Hillel, 1993; Koriat & Goldsmith, 1996). These finding were obtained despite the change in the nature of the test from 4AFC used in Higham and Arnold's research to the T/F format used here. Further, the confidence data demonstrated that self-reported confidence was a predictor of accuracy for the reported items: Correct reported answers were given a significantly higher confidence rating than incorrect re- ported answers.

The present study also showed that providing feedback about optimal bias and instructing participants to distribute their answers between the "Go for points" and "Guess" categories accordingly led to a significant improvement in their test scores. Interestingly, the resolution data demonstrated that the optimal-bias feedback led to a higher increase in the FAR than the HR, which may suggest one reason why examinees tend to be too conservative; specifically, they may have some intuition that responding at optimal bias rather than at their prefeedback bias had forced them to report a considerably larger proportion of their errors (i.e., they are particularly sensitive to their FAR). Despite the large effect of feedback on the FAR, participants moved more correct answers (*M* = 2.99, *SEM* = .37) than incorrect answers (*M* =

1.85, *SEM* = .27) to the “go for points” category in response to optimal-bias feedback, which explains why their corrected score increased.

Although participants moved more correct answer to the “go for points” category in response to optimal-bias feedback, the careful reader may wonder how participants’ corrected score could in- crease after redistributing their responses in response to optimal- bias feedback if the redistribution caused a greater increase to the FAR than the HR; should that not mean that the penalty would overshadow any points gained, leading to a reduction of the corrected score rather than an increase? The reason that there is an increase in the corrected score highlights again how the corrected score is a complex measure made up of several components; in particular, because *f* is above .50, a change to the *b* frequency in Table 1 (false alarms) will have a greater impact on the FAR (*b*/[*b* + *d*]) than a comparable or even larger change to the *a* frequency (hits) will have on the HR (*a*/[*a* + *c*]). In other words, because there are fewer incorrect (*b* + *d*) than correct (*a* + *c*) responses, the FAR is more susceptible to change than the HR.

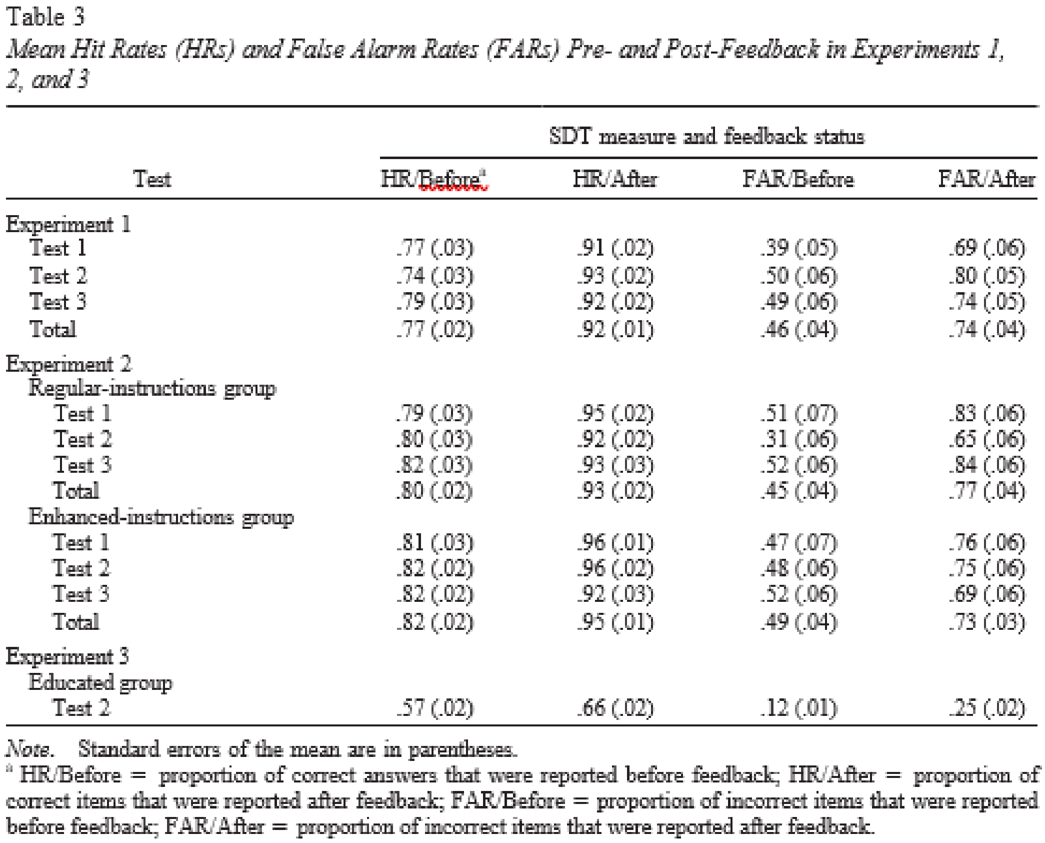

Table 3
Mean Hit Rates (HRs) and False Alarm Rates (FARs) Pre- and Post-Feedback in Experiments 1, 2, and 3

Test	SDT measure and feedback status			
	HR/Before[a]	HR/After	FAR/Before	FAR/After
Experiment 1				
Test 1	.77 (.03)	.91 (.02)	.39 (.05)	.69 (.06)
Test 2	.74 (.03)	.93 (.02)	.50 (.06)	.80 (.05)
Test 3	.79 (.03)	.92 (.02)	.49 (.06)	.74 (.05)
Total	.77 (.02)	.92 (.01)	.46 (.04)	.74 (.04)
Experiment 2				
Regular-instructions group				
Test 1	.79 (.03)	.95 (.02)	.51 (.07)	.83 (.06)
Test 2	.80 (.03)	.92 (.02)	.31 (.06)	.65 (.06)
Test 3	.82 (.03)	.93 (.03)	.52 (.06)	.84 (.06)
Total	.80 (.02)	.93 (.02)	.45 (.04)	.77 (.04)
Enhanced-instructions group				
Test 1	.81 (.03)	.96 (.01)	.47 (.07)	.76 (.06)
Test 2	.82 (.02)	.96 (.02)	.48 (.06)	.75 (.06)
Test 3	.82 (.02)	.92 (.03)	.52 (.06)	.69 (.06)
Total	.82 (.02)	.95 (.01)	.49 (.04)	.73 (.03)
Experiment 3				
Educated group				
Test 2	.57 (.02)	.66 (.02)	.12 (.01)	.25 (.02)

Note. Standard errors of the mean are in parentheses.
[a] HR/Before = proportion of correct answers that were reported before feedback; HR/After = proportion of correct items that were reported after feedback; FAR/Before = proportion of incorrect items that were reported before feedback; FAR/After = proportion of incorrect items that were reported after feedback.

Most importantly, though, the data demonstrated that participants’ corrected score at optimal bias (corrected-feedback) also exceeded their corrected score at forced report (corrected-forced). This finding counters claims in the literature that students writing formula-scored tests should simply be instructed to answer all the questions (e.g., Budescu & Bar-Hillel, 1993). Instead, there is sometimes an optimal level of omissions that is above zero as the SDT model predicts. In Experiment 2, we examine whether these results extend to tests with two plausible alternatives, or whether they are unique to T/F tests.

Experiment 2

Although optimal-bias feedback led to higher scores on each general knowledge test in Experiment 1, there was no tendency for participants to home in on optimal bias with repeated testing. That is, by the final test, participants were still too conservative and were no closer to performing at optimal bias than they had been on Test 1. However, it is possible that the nature of the tests them- selves (i.e., T/F questions) limited the benefit of the optimal-bias feedback; for example, research has shown that performance on T/F tests tends to be significantly lower than two-alternative- forced-choice (2AFC) tests (Jang, Wixted, & Huber, 2009). One limitation of T/F questions is that, in situations where individuals do not know the answer, they cannot use the alternatives (i.e., "true" and "false") to infer the correct response, and therefore they may believe that it is best to remain with a conservative responding strategy. However, when the questions have two informative alternatives rather than just "true" and "false," the alternatives them- selves may provide *partial knowledge* that can be used to infer the correct response. For example, participants may not know the answer to the question "What is the last letter of the Greek alphabet?" but if they were given the two possible alternatives of "Omega" and "Beta" they may remember that "Beta" is the second letter of the alphabet and thus infer that the correct response must be "Omega."

To explore whether the T/F question format hindered the effectiveness of the optimal-bias feedback (such that learning across tests was limited), the T/F statements from Experiment 1 were amended in Experiment 2 such that they were questions with two alternatives. Further, to increase the chances of finding an overall effect of feedback on bias (i.e., a reduction between actual and optimal bias by the final test), half of the participants were given enhanced optimal-bias feedback regarding how performing at optimal bias would have a beneficial effect on their test scores. In particular, in an effort to boost their motivation to take the feed- back seriously, participants in the enhanced-instructions group were shown the score that they could achieve by redistributing their answers between the "go-for-points" and "guess" categories.

Method

Participants. Forty-eight University of Southampton under- graduates participated in exchange for a £5 payment.

Design and materials. The T/F statements from Experiment 1 were edited so that they were questions with two alternative answers (e.g., "What is the last letter of the Greek alphabet?" with "Omega" and "Beta" as the correct and incorrect alternatives, respectively). Additional general-knowledge questions, also edited so that they had two-alternatives, were taken from various sources (e.g., Nelson & Narens, 1980). All three tests contained 20 questions, and question sets were constructed to counterbalance items

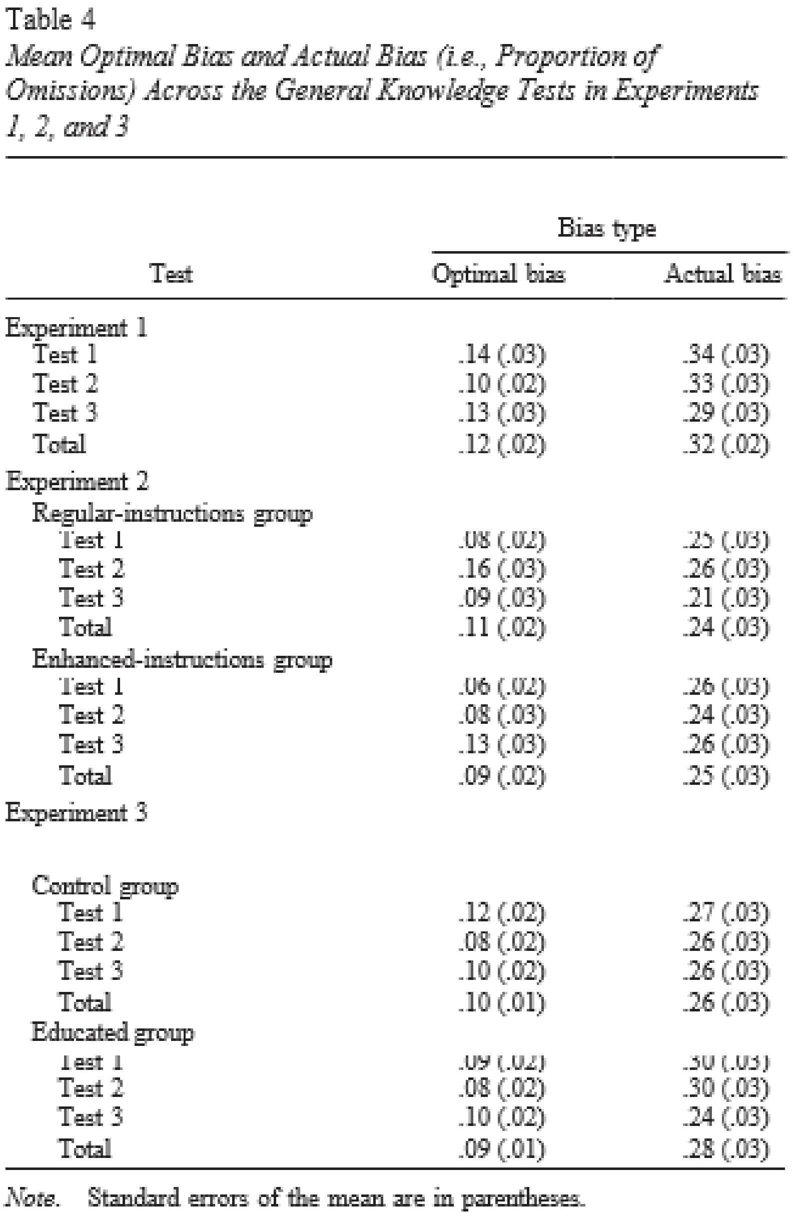

Table 4

Mean Optimal Bias and Actual Bias (i.e., Proportion of Omissions) Across the General Knowledge Tests in Experiments 1, 2, and 3

Test	Bias type	
	Optimal bias	Actual bias
Experiment 1		
Test 1	.14 (.03)	.34 (.03)
Test 2	.10 (.02)	.33 (.03)
Test 3	.13 (.03)	.29 (.03)
Total	.12 (.02)	.32 (.02)
Experiment 2		
Regular-instructions group		
Test 1	.08 (.02)	.25 (.03)
Test 2	.16 (.03)	.26 (.03)
Test 3	.09 (.03)	.21 (.03)
Total	.11 (.02)	.24 (.03)
Enhanced-instructions group		
Test 1	.06 (.02)	.26 (.03)
Test 2	.08 (.03)	.24 (.03)
Test 3	.13 (.03)	.26 (.03)
Total	.09 (.02)	.25 (.03)
Experiment 3		
Control group		
Test 1	.12 (.02)	.27 (.03)
Test 2	.08 (.02)	.26 (.03)
Test 3	.10 (.02)	.26 (.03)
Total	.10 (.01)	.26 (.03)
Educated group		
Test 1	.09 (.02)	.30 (.03)
Test 2	.08 (.02)	.30 (.03)
Test 3	.10 (.02)	.24 (.03)
Total	.09 (.01)	.28 (.03)

Note. Standard errors of the mean are in parentheses.

across the tests, which resulted in each set appearing on each test equally often across participants. Finally, presentation of the questions and tests was identical to Experiment 1 except for three key changes: (1) each question was presented with the correct answer and a plausible foil (randomly assigned to the left or right column on the screen) instead of “true” and “false,” (2) participants were required to report confidence from 50% (no confidence) to 100% (very confident), a confidence scale (unlike the 1– 6 scale used in Experiment 1) that incorporated a blind-guessing baseline, and (3) participants were assigned randomly to one of the two between- subjects conditions of the experiment: *regular-instructions* or *enhanced-instructions*. As in Experiment 1, for responses assigned to the “Go for points” category, the typical correction factor (penalty for errors) was applied (i.e., 1/[number of alternatives — 1]), which for a test with two alternatives, was equal to 1.

Procedure. Participants in the regular-instructions group received the same feedback and information as participants in Experiment 1. However, participants in the enhanced-instructions group were given additional feedback and instructions at the end of each test. Specifically, all participants received the same optimal- bias feedback from Experiment 1, but participants in the enhanced- instructions group were also given information regarding maximum-corrected score; that is, after receiving the optimal-bias feedback based on their performance, the enhanced-instructions group were also shown their predicted maximum-corrected score (i.e., the corrected score they would theoretically obtain if they performed at optimal bias) and were told that they could achieve this score if they correctly adjusted their "go-for-points" versus "guess" assignments.

Results and Discussion

Corrected-score differences. The four different corrected scores across the three tests are presented in Table 2. A 2 (*group*: regular, enhanced) × 3 (*test*: Test 1, Test 2, Test 3) × 4 (*score*: maximum-corrected, corrected, corrected-feedback, corrected- forced) repeated-measures ANOVA demonstrated a main effect of score, *F*(3, 132) = 34.21, *MSE* = .004, h^2 = .44. The planned follow-up comparisons showed that the maximum-corrected score (*M* = .61, *SEM* = .03) was significantly higher than the corrected score (*M* = .54, *SEM* = .02), *t*(45) = 7.23, *p* < .001, and the corrected-forced score (*M* = .58, *SEM* = .03), *t*(45) = 4.79, *p* < .001, but not the corrected-feedback score (*M* = .61, *SEM* = .03), *t*(45) = .65, *p* = .52. As in Experiment 1, the corrected-feedback score was significantly greater than both the corrected score, *t*(45) = 8.31, *p* < .001, and the corrected-forced score, *t*(45) = 4.63, *p* < .001. Finally, the corrected-forced score also was higher than the corrected score, *t*(45) = 4.31, *p* < .001.

There was an interaction between group and test, *F*(2, 88) = 5.30, *MSE* = .06, h^2 = .11, which occurred because there was no difference in the test scores for the regular-instructions group, |*t*|s ≤ 1.86, *p*s ≥ .08, but there was a difference for the enhanced- instructions group. Specifically, the scores in the enhanced- instructions group were lower on Test 3 (*M* = .54, *SEM* = .04) than on either Test 1 (*M* = .62, *SEM* = .03), *t*(23) = 2.22, *p* = .02, or Test 2 (*M* = .61, *SEM* = .04), *t*(23) = 2.11, *p* = .05. One potential interpretation of the lower scores on Test 3 for the enhanced-instructions group is that the added instructions—that is, providing participants with information regarding their maximum- corrected score—negatively impacted their performance. How- ever, inspection of the means for the educated group in Table 2 indicates that, although scores were lower on Test 3, the pattern of the scores is similar to Test 1 and Test 2 (e.g., corrected-feed- back > corrected-forced). Indeed, it appears that the participants in the educated group had a more difficult time overall with Test 3 than the control group (i.e., the corrected-forced score on Test 3, which removes the strategic regulation component, is lower than on Test 1 or Test 2), but that the pattern of their scores on Test 3 replicates the pattern of their previous two tests. There were no other main effects or interactions, *F*s ≤ 1.32, *p*s ≥ .24.

Resolution. To investigate resolution, a 2 (*feedback*: before, after) × 2 (*response*: HR, FAR) × 3 (*test*: Test 1, Test 2, Test 3) within-subjects ANOVA was performed on the proportion of responses placed in the "go for points" column (see Table 3 for the means). There was a main effect both for feedback, *F*(1, 45) = 86.63, *MSE* = .07, h^2 = .65, *p* < .001, and response, *F*(1, 45) = 201.71, *MSE* = .05, h^2 = .82, *p* < .001. The main effect of feedback arose because participants were too conservative pre- feedback; the HRs and FARs were lower before (*M* = .64, *SEM* =

.02) than after (*M* = .85, *SEM* = .02) the feedback was provided. As expected, the main effect of response occurred because the HR (*M* = .87, *SEM* = .01) was higher than the FAR rate (*M* = .61, *SEM* = .02), showing that resolution was reasonably good. As in Experiment 1, there also was an interaction between feedback and

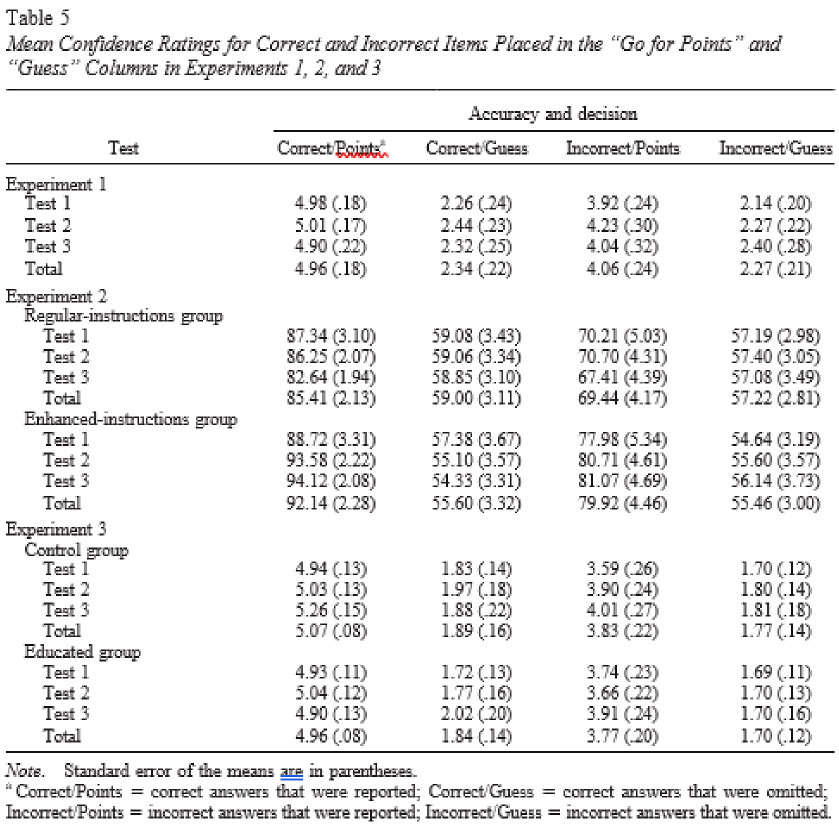

Table 5

Mean Confidence Ratings for Correct and Incorrect Items Placed in the "Go for Points" and "Guess" Columns in Experiments 1, 2, and 3

	Accuracy and decision			
Test	Correct/Points[a]	Correct/Guess	Incorrect/Points	Incorrect/Guess
Experiment 1				
Test 1	4.98 (.18)	2.26 (.24)	3.92 (.24)	2.14 (.20)
Test 2	5.01 (.17)	2.44 (.23)	4.23 (.30)	2.27 (.22)
Test 3	4.90 (.22)	2.32 (.25)	4.04 (.32)	2.40 (.28)
Total	4.96 (.18)	2.34 (.22)	4.06 (.24)	2.27 (.21)
Experiment 2				
Regular-instructions group				
Test 1	87.34 (3.10)	59.08 (3.43)	70.21 (5.03)	57.19 (2.98)
Test 2	86.25 (2.07)	59.06 (3.34)	70.70 (4.31)	57.40 (3.05)
Test 3	82.64 (1.94)	58.85 (3.10)	67.41 (4.39)	57.08 (3.49)
Total	85.41 (2.13)	59.00 (3.11)	69.44 (4.17)	57.22 (2.81)
Enhanced-instructions group				
Test 1	88.72 (3.31)	57.38 (3.67)	77.98 (5.34)	54.64 (3.19)
Test 2	93.58 (2.22)	55.10 (3.57)	80.71 (4.61)	55.60 (3.57)
Test 3	94.12 (2.08)	54.33 (3.31)	81.07 (4.69)	56.14 (3.73)
Total	92.14 (2.28)	55.60 (3.32)	79.92 (4.46)	55.46 (3.00)
Experiment 3				
Control group				
Test 1	4.94 (.13)	1.83 (.14)	3.59 (.26)	1.70 (.12)
Test 2	5.03 (.13)	1.97 (.18)	3.90 (.24)	1.80 (.14)
Test 3	5.26 (.15)	1.88 (.22)	4.01 (.27)	1.81 (.18)
Total	5.07 (.08)	1.89 (.16)	3.83 (.22)	1.77 (.14)
Educated group				
Test 1	4.93 (.11)	1.72 (.13)	3.74 (.23)	1.69 (.11)
Test 2	5.04 (.12)	1.77 (.16)	3.66 (.22)	1.70 (.13)
Test 3	4.90 (.13)	2.02 (.20)	3.91 (.24)	1.70 (.16)
Total	4.96 (.08)	1.84 (.14)	3.77 (.20)	1.70 (.12)

Note. Standard error of the means are in parentheses.

[a] Correct/Points = correct answers that were reported; Correct/Guess = correct answers that were omitted; Incorrect/Points = incorrect answers that were reported; Incorrect/Guess = incorrect answers that were omitted.

response, *F*(1, 45) = 26.54,$_p$ *MSE* = .03, h^2 = .37, *p* < .001, which occurred because the difference between pre- and postfeedback for the HRs versus FARs showed that feedback led to a larger increase in the FAR (*M*diff = .28, *SEM* = .04) than the HR (*M*diff = .13, *SEM* = .01). No other main effect or interactions were significant, *F*s ≤ 2.09, *p*s ≥ .13.

$_p$ **Optimal versus actual bias.** Bias profiles were created for each group of participants for the three tests, and optimal and actual bias are reported in Table 4. These data were analyzed in a 2 (*bias type*: optimal, actual) × 2 (*group*: regular, enhanced) × 3 (*test*: Test 1, Test 2, Test 3) mixed ANOVA, with group as the between-subjects factor. The analysis revealed a main effect of bias type, *F*(1, 46) = 78.60, *MSE* = .02, h^2 = .63, *p* < .001, and an interaction between test and group, *F*(2, 92) = 5.23, *MSE*^{p}= .01, h^2 = .10, *p* = .01. No other main effect or interaction was significant, *F*s ≤ 2.30, *p*s ≥ .11.

The main effect of bias type arose because participants were too conservative in their responding; across all three tests, actual bias (*M* = .24, *SEM* = .02) was significantly higher than optimal bias (*M* = .10, *SEM* = .01). Follow-up *t* tests for the test by group interaction demonstrated that there was a different pattern of average bias scores between Test 2 and 3 for the regular- and enhanced-instructions groups; that is, the bias scores in the regular-instructions group were higher on Test 2 (*M* = .21, *SEM* = .03) than on Test 3 (*M* = .15, *SEM* = .02), *t*(23) = 2.07, *p* = .05, whereas the reverse pattern was observed in the enhanced-instructions group (Test 2: *M* = .16, *SEM* = .02; Test 3: *M* = .19, *SEM* = .02), *t*(23) = 2.52, *p* = .02. The average bias scores also were lower on Test 1 (*M* = .16, *SEM* = .02) than Test 3 for the enhanced-instructions group, *t*(23) = 2.45, *p* = .02 but the com- parable difference in the regular-instructions group was not significant, *t*(23) = .49, *p* = .63. No other significant differences were found from the follow-up comparisons, |*t*|s ≤ 1.94, *p*s ≥ .07. **Confidence ratings.** The group variable (regular vs. enhanced) was not included in the analysis because of insufficient sample size. Specifically, not all participants had H, M, FA, and CR responses in every test, which is not an issue for the preceding analyses (e.g., a FAR of 0 simply means the corrected score is equal to the HR), but does lead to problems for analyzing confidence if both group and test are included as variables. Therefore, only participants who contributed data to all cells across the three tests were included in the analysis, which reduced the group size to eight participants in the regular-instructions condition and seven participants in the enhanced-instructions condition. The mean confidence ratings for both groups are shown in Table 5. A 2 (*response*: correct, incorrect) × 2 (*report decision*: go for points, guess) × 3 (*test*: Test 1, Test 2, Test 3) mixed ANOVA was conducted on the mean ratings. The results showed main effects of response, *F*(1, 14) = 46.57, *MSE* = 56.06, h^2 = .77, *p* < .001, and report decision, *F*(1, 14) = 91.79, *MSE* = 265.08, h^2 = .87, *p* <.001, but not of test, *F* < 1. The main effects of response and report decision occurred because, as expected, participants as- signed higher confidence ratings to correct (*M* = 72.98, *SEM* = 1.59) than incorrect (*M* = 65.37, *SEM* = 2.18) responses, and to responses placed in the “go for points” (*M* = 81.44, *SEM* = 2.36) than “guess” (*M* = 56.91, *SEM* = 2.09) column. There also was an interaction between response and report decision, *F*(1, 14) = 29.42, *MSE* = 66.69, h^2 = .68, *p* < .001, which arose because the difference in confidence ratings between correct and incorrect responses was larger for items placed in the “go for points” column (*M*diff = 14.22, *SEM* = 2.24) than for the items placed in the “guess” column (*M*diff = 1.01, *SEM* = .67). No other interactions were significant, all *F*s ≤ 1.

Summary. The results from Experiment 2 replicated the pat- terns found in Experiment 1; specifically, participants were under- confident in their knowledge and, although optimal-bias feedback and answer redistribution improved performance on every test, there was no strong reliable reduction between actual and optimal bias by the third general knowledge test. Further, the confidence data again showed that participants’ self-reported confidence was a predictor of accuracy for the reported items; that is, correct reported answers were given a significantly higher confidence rating than incorrect reported answers. Additionally, the resolution data demonstrated that the optimal-bias feedback led to a higher increase in the FAR than the HR. Again though, the frequency data showed that participants did move more correct answers (*M* = 2.16, *SEM* = .21) than incorrect answers (*M* = .88, *SEM* = .13) to the “go for points” category after receiving optimal-bias feed- back.

Importantly, as in the previous experiment, the data from Experiment 2 demonstrated that performance at optimal bias led to a significantly higher corrected test score than if participants simply had been told to report all answers (corrected-feedback > corrected-forced). However, enhancing the optimal-bias feedback to include information regarding maximum-corrected score (i.e., the enhanced-instructions condition) did not significantly improve performance over the regular feedback. Further, switching from the T/F format used in Experiment 1 to two-alternatives in the present experiment had almost no impact on the data patterns. For example, the optimal-bias feedback was not more effective for long-term learning with two-alternative questions than T/F questions: As in Experiment 1, participants' actual bias was no closer to their optimal bias by the final test.

Indeed, the only small difference in the data for the present study compared to Experiment 1 was that the maximum-corrected score was not significantly higher than the corrected-feedback score. This difference likely was due to the change in the question format, as previous research has found that accuracy tends to be lower with T/F questions (e.g., Jang et al., 2009), and Table 3 suggests that participants in Experiment 1 may have had more difficulty with resolution (i.e., a smaller difference between the HR and FAR in Experiment 1 compared to Experiment 2). Specifically, there is less information in the T/F answer choices than the two-alternatives answer choices to make the decision on which items to move over from the "guess" column to conform to the optimal-bias feedback. This lack of information may have affected participants' motivation to perform the task as requested in Experiment 1, potentially introducing a random component to their decisions and causing the corrected-feedback score to fall short of the maximum-corrected score.

Experiment 3

In both Experiment 1 and 2 participants were required to apply what they had been told about their optimal bias to the test on which it had been computed; that is, on each test their optimal bias was calculated and subsequently they were required to review the same test and adjust the number of responses in the go-for-points/ guess columns accordingly. Experiment 3 was designed to test the efficacy of feedback when optimal-bias scores from an initial test are applied to performance on a *different* test.

There are two main reasons why it is important to explore whether optimal-bias feedback transfers to different tests. First, it is possible that no overall reduction in actual bias from the first to final general knowledge test was found in Experiments 1 and 2 because participants may not have understood that the optimal-bias score on each test was telling them something about their general responding strategies. More specifically, because feedback was given on a test-by-test basis, participants may not have sufficiently internalized that overall they were too conservative (i.e., that independent of any given test, they "guess" too many items). By having participants use previous test performance to adjust their criterion on a current test, the generality of the problem may become more apparent.

Second, using optimal-bias feedback to improve performance is less effective if the feedback does not transfer across different tests. That is, individuals who want to improve their performance on standardized tests such as the SAT Reasoning Test will need to complete the training before taking the actual test itself, and therefore the benefit from this training needs to carry-over from practice tests. Because large-scale, formula-scored tests such as the SAT Reasoning Test have multiple alternatives rather than just two as in Experiments 1 and 2, multiple-choice questions with four alternatives were used in Experiment 3. Presenting multiple alter- natives per question also allowed us to more directly compare the current results obtained in a controlled experimental context to Higham and Arnold's (2007a) classroom data.

Method

Participants. Forty-eight University of Southampton under- graduates participated in exchange for course credit in an introductory psychology course or a £5 payment.

Design and materials. The three general knowledge tests contained questions used in Experiment 1 and 2; there were 50 questions on both Test 1 and Test 2, and 35 questions on Test 3. As in Experiment 2, the third test contained a different number of questions to dissuade students from simply mimicking postfeed- back performance (i.e., "guessing" the same number of times on Test 3 as on Test 2 that received the optimal-bias feedback).

Each question was presented with four possible alternatives, which consisted of the correct answer plus three plausible incorrect responses. Unlike the previous two experiments, participants in this experiment were given paper/pencil tests instead of answering the questions on a computer. The answer sheet for the test questions included the following (1) a section containing the four possible response alternatives (labeled a, b, c, and d), (2) a section

with two separate columns to reflect the point system, with one column designated "Go for points" and the other column designated "Guess," and (3) a column to record confidence in the chosen response on a scale from 1 to 6 (with 1 being least confident and 6 being most confident). For responses assigned to the "Go for points" category, the typical correction factor (penalty for errors) was applied (i.e., 1/[number of alternatives — 1]), which for a test with four alternatives, was equal to 0.33. For each response category, the points gained or lost for correct and incorrect answers was indicated on the answer sheet (i.e., "+1/—0.33" for "Go for points" and "0 points" for "Guess").

Three sets of 50 questions (question set: A, B, and C) were constructed to counterbalance the items across the tests. To create the 35 question set versions for Test 3, 15 questions from each set were deleted randomly, and the same 15 items from each set were always omitted from Test 3. The questions sets were counterbalanced so that each set appeared equally often across participants in Test 1, 2, or 3. Participants were assigned randomly to one of the two between-subjects conditions of the experiment; *control* or *educated*. All participants completed the three general knowledge tests, and the only difference between the two groups was the content of the instructions given on Test 2.

Procedure. Participants were tested individually or in groups of no more than three people. To ensure that participants under- stood the structure of the tests, they were given a set of instructions to read prior to starting Test 1, which were verbally reiterated by the experimenter. Both the written instructions and the experimenter informed the participants that they had two options when answering a question; that is, after circling a response on the test sheet they had to choose whether or not they wanted that response to count in the point tally. More specifically, they were informed that if they believed that they knew the correct answer to a question then they should mark the "Go for points" column—if they circled the correct answer they would receive 1 point, but if they gave an incorrect response they would be penalized .33 points. Alternatively, if they believed that they did not know the correct answer to a question (i.e., that they were guessing) then they were told to mark the "Guess" column, in which case no points would be gained or lost. Finally, participants were told that, regardless of whether they chose to "go for points" or "guess," they had to rate their confidence in their chosen response using the scale of 1 to 6.

At the end of Test 1 all participants were given a 6-minute embedded figures task. This filler task allowed the experimenter to score the first test and compute each participant's optimal-bias score (i.e., the number of responses that they should have placed in the "guess" column to achieve the maximum-corrected score). Test 2 was administered after the embedded figures task, and all participants were given the same instructions as Test 1. After completing Test 2 the participants in the control condition were given Test 3, with the only added instruction that this test contained 35 instead of 50 questions. Conversely, participants in the educated group were given optimal-bias feedback. That is, the educated group was told that their optimal-bias scores had been calculated on Test 1 and that they were restricted to this number on Test 2; therefore, if performance was not at optimal bias, they were required to go back over the items on Test 2 that they had placed in the "guess" and "go for points" column and redistribute them accordingly. As in the previous experiments, the participants were told that they were not allowed to make any changes to their actual responses or their confidence ratings. Once participants in the educated condition had the required amount of responses placed in the "Guess" and "Go for points" columns (i.e., had achieved their Test 1 optimal bias) they were allowed to move on to Test 3.

Results

Preliminary analyses. Because we used performance on Test 1, which used one question set, to guide performance on Test 2, which used a different question set, it was important to demonstrate similar levels of optimal bias between the three question sets. A 2 (*test*: Test 1, Test 2) × 6 (*set/order*: AB, AC, BA, BC, CA, CB) mixed ANOVA was performed on the optimal-bias scores, with set/order as the between-subjects variable. If optimal bias varied systematically between question sets, this variation would emerge from this analysis as an interaction between test and set/order. There were no main effects of test, $F < 1$, or set/order, $F(5, 61) = 2.15$, $p = .07$. More importantly, there was no inter- action between test and set/order, $F(5, 61) = 1.88$, $p = .11$.

Corrected-score differences. The different corrected scores across the three tests are presented in Table 2. Unlike previous experiments, the first analysis excluded the corrected-feedback score because feedback was only provided to participants in the educated group for the second test. A 2 (group: control, educated) × 3 (score: maximum-corrected, corrected, corrected- forced) × 3 (test: Test 1, Test 2, Test 3) mixed ANOVA was conducted, with group as the between-subjects variable. The results revealed a main effect of score, $F(2, 92) = 102.41$, *MSE* =

.002, $h^2 = .69$, $p < .001$, but no other main effect or interaction was significant, $F\text{s} \leq 1.33$, $p\text{s} \geq .27$. Follow-up *t* tests demonstrated that the main effect of score occurred because the maximum-corrected score ($M = .61$, $SEM = .02$) was significantly higher than both the corrected score ($M = .55$, $SEM = .02$), $t(47) = 12.16$, $p < .001$, and the corrected-forced score ($M = .59$, $SEM = .02$), $t(47) = 8.58$, $p < .001$. Additionally, the corrected- forced score also was significantly higher than the corrected score, $t(47) = 8.29$, $p < .001$.

Planned-comparisons for Test 2 of the educated group (i.e., where feedback based on Test 1 optimal bias occurred) were conducted to compare the four different corrected scores. The maximum-corrected score was significantly higher than the corrected score, $t(23) = 6.85$, $p < .001$, the corrected-feedback score, $t(23) = 5.43$, $p < .001$, and the corrected-forced score, $t(23) = 5.16$, $p < .001$. Importantly, the corrected-feedback score was significantly higher than the corrected score, *t*(23) = 4.53, *p* <.001. The corrected-forced score was also higher than the corrected score, *t*(23) = 5.00, *p* < .001; however, there was no significant difference between the corrected-feedback and corrected-forced scores, *t*(23) = 1.89, *p* = .07.

Resolution. To explore resolution on Test 2 of the educated group, a 2 (feedback: before, after) × 2 (response: HR, FAR) within-subjects ANOVA was performed on the proportion of responses placed in the "go for points" column (see Table 3). There was a main effect both for feedback, *F*(1, 23) = 80.49, *MSE* = .003, h^2 = .78, *p* < .001, and response, *F*(1, 23) = 219.48, *MSE* = .02, h^2 = .91, *p* < .001. The main effect of feedback arose because every participant in the educated group was too conservative prefeedback; the proportion of responses assigned to the "Go for points" category was lower before (*M* = .35, *SEM* = .01) than after (*M* = .46, *SEM* = .01) the feedback occurred on Test 2. As anticipated, the main effect of response occurred because the HR (*M* = .62, *SEM* = .02) was higher than the FAR (*M* = .19, *SEM* = .01), showing that resolution was reasonably good. Of more interest, though, was the interaction between feedback and response, *F*(1, 23) = 9.04, *MSE* = .001, h^2 = .28, *p* = .006. As in the previous two experiments, the difference between pre- and postfeedback for the HRs versus FARs showed that feedback led to a larger increase in the FAR (*M*diff = .13, *SEM* = .01) than for the HR (*M*diff = .09, *SEM* = .01).

Optimal versus actual bias. Bias profiles were created for each group of participants for the three tests, and optimal and actual bias are reported in Table 4. These data were analyzed in a 2 (*bias type*: optimal, actual) × 2 (*group*: control, educated) × 3 (*test*: Test 1, Test 2, Test 3) mixed ANOVA, with group as the between-subjects factor. The analysis revealed a main effect of bias type, *F*(1, 46) = 111.65, *MSE* = .02, h^2 = .71, *p* < .001, and a bias type by test interaction, *F*(2, 92) = 3.66, *MSE* = .005, h^2 =.07, *p* = .03. The main effect of bias type arose because participants were too conservative in their responding; across all three tests, actual bias (*M* = .27, *SEM* = .02) was significantly higher than optimal bias (*M* = .09, *SEM* = .01).

To explore the interaction, difference scores between the bias score types were calculated, and follow-up *t* tests indicated that the difference between the two bias scores was smaller on Test 3 (*M*diff = .15, *SEM* = .02) than on Test 2 (*M*diff = .20, *SEM* = .02), *t*(47) = 2.79, *p* = .008. There was no difference in the bias score differences between Test 1 (*M*diff = .18, *SEM* = .02) and Test 2, *t*(47) = 1.11, *p* = .27, nor between Test 1 and Test 3, *t*(47) = 1.44, *p* = .16. As can be seen from Table 4, performance across the tests was relatively stable for participants in both the control and educated groups. However, the means for the educated group showed a reduction in omitted responses (i.e., actual bias) on the final test. These observations were confirmed in planned follow-up analyses. That is, there was no difference across the tests for amount of withheld responses in the control group, all |*t*|s < 1, but participants in the educated group withheld less on Test 3 than they did on either Test 1, *t*(23) = 3.10, *p* = .005, or Test 2, *t*(23) = 3.69, *p* < .001.

Confidence ratings. The mean confidence ratings for both groups are shown in Table 5. A 2 (*group*: control, educated) × 2 (*response*: correct, incorrect) × 2 (*report decision*: go for points, guess) × 3 (*test*: Test 1, Test 2, Test 3) mixed ANOVA was conducted on the mean ratings, with group as the between-subjects factor. This analysis was performed on the prefeedback Test 2 data for the educated group. The results showed no main effect of group, *F* < 1, but there were main effects of response, *F*(1, 38) = 124.34, *MSE* = .43, h^2 = .77, *p* < .001, report decision, *F*(1,38) = 1232.21, *MSE* = .66, h^2 = .97, *p* < .001, and test, *F*(2,

76) = 4.01, *MSE* = .27, h^2 = .10, *p* = .02. The main effects of response and report decision occurred because, as expected, participants assigned higher confidence ratings to correct (*M* = 3.44, *SEM* = .07) than incorrect (*M* = 2.77, *SEM* = .11) responses, and to responses placed in the "go for points" (*M* = 4.41, *SEM* = .10) than the "guess" (*M* = 1.80, *SEM* = .10) column. The main effect of test was due to an increase in confidence ratings from Test 1 to Test 3; specifically, confidence ratings were higher on Test 3 (*M* = 3.18, *SEM* = .10) than Test 1 (*M* = 3.02, *SEM* = .09), *t*(40) = 2.36, *p* = .02. However, the increase from Test 1 to Test 2 (*M* = 3.11, *SEM* = .10), and Test 2 to Test 3, failed to reach significance, both |*t*/s ≤ 1.88, *p*s ≥ .07.

There also was an interaction between response and report decision, *F*(1, 38) = 63.45, *MSE* = .55, h^2 = .63, *p* < .001, which occurred because the difference in confidence ratings between correct and incorrect responses was larger for items placed in the "go for points" column (*M*diff = 1.21, *SEM* = .12) than for the items placed in the "guess" column (*M*diff = .13, *SEM* = .05). No other interactions were significant, all *F*s ≤ 2.43, *p*s ≥ .10.

Summary. The data from Experiment 3 demonstrated that providing feedback to participants—that is, using the optimal-bias score calculated on one test to guide test-taking strategy on a subsequent test— does improve performance. Specifically, the participants who were informed of their Test 1 optimal bias and required to apply this information to their Test 2 performance (i.e., by adjusting the number of items they had placed in the "guess" and "go for points" columns accordingly) showed a significantly higher corrected score postfeedback. However, unlike Experiments 1 and 2, the corrected score increase attributable to optimal-bias feedback did not exceed the corrected score at forced report. This issue will be addressed further in the General Discussion.

Feedback led to a significant increase in the prefeedback corrected scores on Test 2 even though participants moved more incorrect than correct responses from the "guess" to "go for points" column postfeedback. That is, unlike in Experiment 1 and 2, the frequency data showed that participants in the educated group moved more incorrect answers (*M* = 6.29, *SEM* = .72) than correct answers (*M* = 4.46, *SEM* = .63) to the "go for points" category on Test 2 after receiving optimal-bias feedback. How- ever, the prefeedback corrected score still increased in response to feedback because the penalty was lower in this experiment (.33) than in the previous ones (1; i.e., the ratio of incorrect to correct responses moved from "guess" to "go for points" was less than 3:1).

The results of the current study also showed that participants in the educated group reduced the proportion of responses they withheld on Test 3 so that it was significantly lower than the proportion they withheld on both Test 1 and Test 2 and closer to optimal bias. This finding demonstrates for the first time that feedback regarding optimal criterion setting provided on one test can carry over to subsequent tests, which is promising if optimal bias training programs are to be developed that have long-term effects. However, the rate of withholding on Test 3 in the educated group was still too high, suggesting that participants were still somewhat reluctant to reduce their rate of withholding even after receiving feedback on Test 2.

Supplemental Pooled Analyses

Ranking

What do increases in performance attributable to feedback on optimal bias mean in real-world terms? Would students who wrote a standardized formula-scored test such as the SAT Reasoning Test perform noticeably better if they performed at optimal bias compared to other students who do not? One way to answer this question is to consider changes in percentile rank that might result from optimal-bias feedback. Standardized test results typically take the form of some kind of ranking statistic computed with respect to the current cohort of examinees. To examine any changes to percentile rank that may have resulted from our optimal-bias feedback, we first computed the percentile rank of each student's corrected score based on the distribution of corrected scores that existed prior to feedback. Then, for each student, a new percentile rank was computed from that student's corrected- feedback score, but that rank was again based on the distribution of prefeedback corrected scores. Finally, we computed a difference score between each student's percentile rank before versus after feedback. Thus, with this analysis, it was possible to determine how much a single examinee's percentile rank would increase in their cohort of no-feedback examinees if only that single individual received feedback and performed at optimal bias instead of at the bias associated with no feedback (i.e., the bias associated with the corrected score, which tends to be overly conservative).

The analysis showed that average percentile rank increased for all 10 tests across all three experiments. In particular, the increases were: (1) +4 percentiles, +11 percentiles, and + 13 percentiles for Tests 1, 2, and 3 of Experiment 1, respectively, (2) +14

percentiles, +8 percentiles, and +12 percentiles, for Tests 1, 2, and 3, of the Regular Instructions group of Experiment 2, respectively, and (3) +24 percentiles, +12 percentiles, and +11 percentiles, for Tests 1, 2, and 3, of the Enhanced Instructions group of Experiment 2, respectively, and (4) +13 percentiles for Test 2 of the Educated group of Experiment 3. The average percentile rank increase averaged across all tests used in the three experiments was + 11 percentiles. In some cases, the change in percentile rank could not be computed because the corrected-feedback score exceeded the maximum value of the (prefeedback) corrected scores. In no cases did the corrected-feedback score fall below the mini- mum of the (prefeedback) corrected scores. This analysis suggests that the small but reliable increase in the corrected score resulting from optimal criterion placement that we have observed in our research is likely to have substantial effects on standardized test performance. Many students would likely be thrilled at the proposition of increasing their ranking by 11 percentiles simply by having been instructed on how often to omit answers.

The preceding analysis assumes that there is a single examinee who receives and conforms to optimal bias training whereas others within the same cohort do not. However, what would ranking results look like if *all* examinees received this training? Also, how would ranking scores compare between a case in which all examinees perform at optimal bias (corrected-feedback score) versus one in which they answer all the questions (corrected-forced score)? Certainly, it is more straightforward to instruct examinees to answer all questions than it is to compute optimal bias and attempt to train students to perform at that omission rate. Hence, if the ranking scores are identical between the two cases, efforts may be better directed at persuading students to avoid omissions altogether than teaching them more complex omission strategies. Although universal conformity to *any* training regime is unlikely to happen for actual, large-scale formula scored tests such as the SAT Reasoning Test, a comparison of ranking statistics may nonetheless be informative about the real-world impact of optimal- bias feedback and how it compares to an answer-all strategy. It should also be informative about what aspects of performance the answer-all versus optimal-bias ranking statistics are actually measuring To explore this issue in more depth, we computed the percentile ranks based on the distribution of corrected-feedback scores (optimal-bias strategy) and the percentile ranks based on the distribution of corrected-forced scores (forced-report strategy) for Experiment 1 (collapsed across tests), the two groups of Experiment 2 (separately for each group and collapsed across tests), and for the second test in the educated group of Experiment 3. We then computed a difference score between the two percentile ranks; positive versus negative values translate into higher ranking at optimal bias versus forced report, respectively. If the corrected- feedback and corrected-forced score rankings were identical, all participants would yield a difference score of zero. However this was not the case: only 26 of 96 (27%) participants had ranking scores that were unchanged. The mean absolute change in percentile ranking within each group was 6.29, (*SEM* = 0.60; see Figure 3, explained in more detail below).

What accounts for the difference in ranking? According to the SDT model, corrected scores are a complex measure involving not just the sheer ability to answer the question correctly (retrieval), but also resolution and bias. However, this is only true if the omission rate is above 0% or below 100% (Higham & Arnold, 2007b), which is not the case at forced report (omission rate = 0%). Hence, only the corrected-feedback score can be affected by these parameters. However, because the corrected-feedback score has, by definition, had the negative influence of a nonoptimal criterion setting eliminated, only resolution is left to account for the difference in ranking.

To test this prediction we examined the relationship between resolution and the percentile ranking difference scores. For this analysis we used *d*= scores at optimal bias as our measure of resolution, and the results are shown in Figure 3, which plots the size of the percentile ranking difference scores for the three experiments as a function of *d*=. Data from 28 participants (across all three experiments) were not plotted because *d*= was undefined. As the figure makes clear, the relationship was positive and strong ($r = .68$, $p < .001$). Specifically, examinees with poorer-than- average resolution had better ranking if their test was scored at forced report (corrected-forced score), presumably because that score eliminates the negative influence of their poor resolution on their score. Conversely, examinees with better-than-average resolution were better off if their test was scored at optimal bias (corrected-feedback score). Clearly, because we observed an over- all increase in the corrected-feedback score relative to the corrected-forced score when all participants were considered (at least in Experiments 1 and 2), most participants must have had sufficient resolution skills to benefit from optimal-bias feedback. Importantly, because the corrected-forced score is just a mono- tonic transformation of number-right scoring (and *f*), the ranking of the corrected-forced score doubles as the ranking for number-right (and *f*) scoring. That is, because the omission rate is fixed at zero and cannot affect any score at forced report, the ranking remains the same at forced report regardless of whether it is based on number right, *f*, or the corrected-forced score. Thus, the same result shown in Figure 3 (and the conclusions derived from it) would obtain if resolution was used to predict the ranking difference between tests scored for number-right (or *f*) versus the same tests formula scored at optimal bias (i.e., corrected-feedback score).

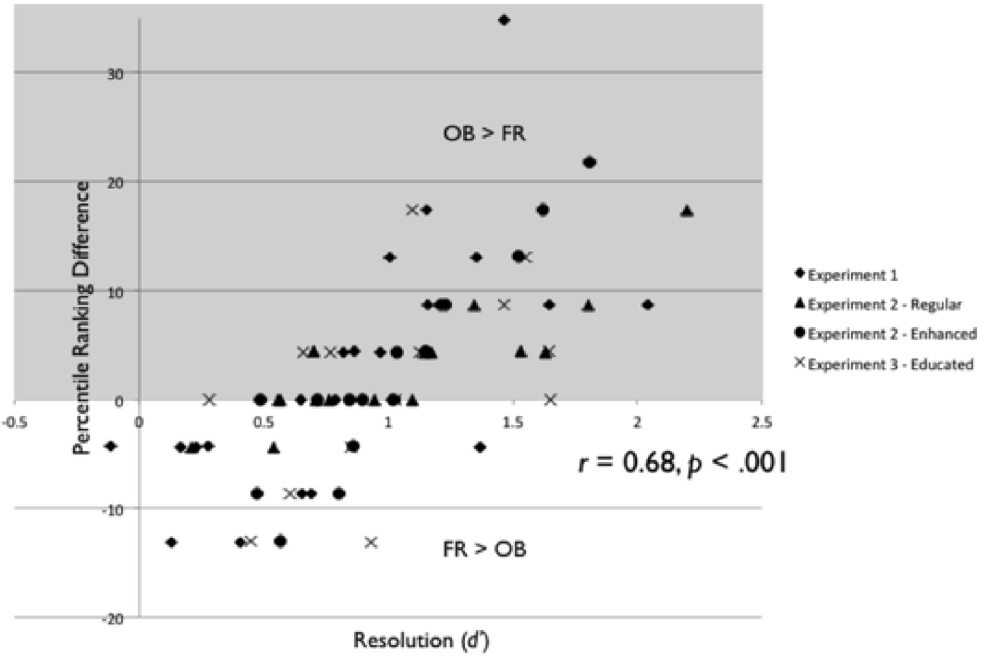

Figure 3. Percentile ranking difference (corrected-score percentile rank at optimal bias [OB] minus the corrected-score percentile rank at forced report [FR]) as a function of resolution (*d*=, computed at OB) for Experiments 1–3. The shaded region shows where OB > FR, whereas the nonshaded region shows where FR > OB. FR is identical to number-right ranking.

Withheld-Item Accuracy

How could it be that forcing output (i.e., reporting all answers) produced a lower corrected-forced score than a corrected-feedback score in Experiments 1 and 2? If the answers omitted at optimal bias were of chance-level (50%) accuracy or higher, then “releasing” them by forcing output should either make no change to the score or improve it (Budescu & Bar-Hillel, 1993). Thus, there is the seemingly paradoxical situation in that, whereas the prefeed- back omitted answers must have had above-chance accuracy for the corrected score to improve at optimal bias, the postfeedback omitted answers must have had below-chance accuracy to cause the score to drop at forced report.

To investigate this issue in more detail, we compared the accuracy of omitted answers in Experiment 2 before and after feedback. In particular, we conducted a 2 (feedback: before, after) × 2 (group: regular, enhanced) mixed ANOVA, with feedback as the repeated-measures factor. We chose to restrict our analysis to Experiment 2 data because (1) the pattern of corrected-forced < corrected-feedback was observed in that experiment, and (2) the inclusion of two experimental groups produced sufficient power to detect differences. The analysis revealed a main effect of feedback, *F*(1, 43) = 13.30, *MSE* = 0.04, h^2 = .24, *p* = .001, which occurred because the accuracy of omitted answers before feedback (*M* = .57, *SEM* = .02) was greater than after feedback (*M* = .41, *SEM* = .02). Also, accuracy of the omitted answers prefeedback was significantly above chance, —95% confidence limit = .52, whereas the accuracy of omitted answers at optimal bias after feedback was significantly below chance, +95% confidence limit = .49.

This analysis suggests an explanation for why increasing output had a nonmonotonic effect on the corrected score: The answers that were moved from the “Guess” category to the “Go for Points” category in response to feedback were of good quality, whereas the answers that were left behind in the “Guess” category at optimal bias after the move was complete were not. In other words, participants were able successfully to choose originally omitted answers to risk against the penalty to improve their score post feedback. This finding is promising news for students who want to eliminate bias effects from their scores on formula-scored exams. One concern about this result is that it is specific to our question sets. For example, it is conceivable that there were some so-called “deceptive” questions in the set with incorrect alternatives that participants consistently endorsed with greater-than-chance likelihood. However, this situation seems unlikely for at least two reasons. First, incorrect answers to deceptive questions are typically endorsed with high confidence (e.g., Koriat & Goldsmith, 1996), whereas the omitted answers that remained after feedback in the current studies were at the lowest level of confidence. Second, the same pattern of results was obtained in Experiment 1, in which there were no potentially deceptive alternatives to con- sider (i.e., only T/F). Nonetheless, we thought it would be prudent to conduct an analysis of the items we used in Experiment 1 and 2 (where all participants received optimal-bias feedback) to explore the possibility of our having deceptive questions in our sample. In particular, we conducted one-sample *t* tests on the accuracy of postfeedback answers in Experiment 1 and 2 that were omitted by at least one participant. The analysis showed that these items as a whole had accuracy rates of 67% (Experiment 1) and 76% (Experiment 2), which were significantly above the chance rate of 50% [Experiment 1: *t*(51) = 10.84, *p* < .001, Experiment 2: *t*(56) = 6.44, *p* < .001]. This overall above-chance accuracy for the items that continued to be withheld after optimal-bias feedback was also found with a more strict analysis, that is, for the items that were *most often* left in the “guess” column postfeedback. For example, the mean accuracy in Experiment 2 of the 10 most frequently omitted questions postfeedback was 65%, which a one-sample *t* test again showed was significantly above chance, *t*(9) = 2.41, *p* = .04.

Together, these analyses show that, although on a per participant basis, the postfeedback omitted answers had below-chance accuracy, this did not come about because participants as a whole were generally deceived by these questions or their alternatives. Instead, it suggests that, for any given individual, there are a subset of answers associated with very low confidence that have a below- chance accuracy rate, but that the items that make up this rate vary between participants.

General Discussion

Similar to our previous results (e.g., Higham, 2007; Higham & Arnold, 2007a), participants in the current set of experiments demonstrated underconfidence. That is, they consistently underes- timated their knowledge, omitting high-quality responses instead of reporting them and thereby suffering an opportunity cost. This result was obtained despite substantive differences in the nature of the test in Experiments 1 and 2 compared with previous research (T/F and two-alternative, rather than four-alternative). Coupled with Higham’s (2007) SAT Reasoning Test results, the current data suggest that underconfidence is a fairly ubiquitous phenom- enon and not specific to the small reward for correct “guesses” (which ostensibly could have attracted responding from risk- averse students) used in Higham and Arnold’s (2007a) research. Further, the results from all three experiments demonstrated that informing participants of their optimal-bias score derived from bias profiles—and having them redistribute their omitted versus reported answers—is a viable technique for helping people adjust their test-taking strategies and improve their performance. Not only did this feedback enhance the corrected score relative to participants' "natural" report criterion, but in Experiments 1 and 2, it also enhanced it relative to forced report (i.e., corrected-feedback score > corrected-forced score). This finding indicates that, contrary to the advice of some theorists (Budescu & Bar-Hillel, 1993), it is not always the best strategy to answer all questions on formula-scored tests; omitting some answers, particularly those associated with very low levels of confidence, is sometimes beneficial.[1]

[1] It is important to keep in mind that the bias profiles created from type-2 SDT measures assume an equal-variance Gaussian (EVG) model of the underlying evidence distributions (i.e., correct vs. incorrect items). Admittedly, if the underlying model is not EVG, bias profiles may be slightly inaccurate in terms of the number of answers to redistribute. However, we were not overly concerned about this issue for the present paper because the information that bias profiles provided was accurate enough to improve students' scores, and *d*= was a good predictor of examinees' change in ranking. Nonetheless, future research should attempt to generalize bias profiles beyond the EVG model scenario to get even more accurate estimates of test-takers performance.

Implications for Withheld-Item Accuracy Results

The analysis conducted on the items used in Experiments 1 and 2 (reported in the "Supplemental Pooled Analyses" section at the end of Experiment 3) demonstrated that there is a subset of items specific to individual participants that have below-chance accu racy. This result is consistent with an SDT model of confidence and accuracy on multiple-choice tests, which allows for the pos- sibility of inverted U-shaped bias profiles (see Figure 2). However, it runs counter to the notion that there is a chance-level accuracy baseline of 1/n corresponding to blind guessing. As Higham (2007) noted, students reason their way to particular answers (whether those answers are correct or not), and they seldom blindly guess. Reassuringly, students seemed to be able to discriminate between correct versus incorrect answers even if their initial decision was to omit them, as long as they were given feedback about their optimal omission rate. The future challenge for students and educators alike is to create situations in which students choose their optimal omission rate for themselves rather than having to be told it, a topic to which we turn next. For now at least, we know that the optimal rate is not always zero (forced report), a finding that is both surprising and contradicts the received wisdom in the decision- making literature.

Although the data from Experiment 1 and 2 showed a difference between the corrected-forced score and the corrected-feedback score, this difference did not replicate in Experiment 3. The likely reason for this is that, all else being equal, the low penalty (p =

.33) used for the four-alternative tests in Experiment 3 produces less bowing in the bias profiles compared to higher penalties (p = 1), such as those for the two-alternative tests in Experiments 1 and 2 (see Figure 6 in Higham & Arnold, 2007b). Stated differently, compared with low penalties, there is a higher sunk cost for reporting too much with large penalties, producing a greater difference between the corrected score at optimal bias and that at forced report.

Implications for Ranking Results

The ranking analyses presented in the “Supplemental Pooled Analyses” section demonstrated that the optimal-bias feedback led to a sizable average percentile rank increase compared to prefeed- back ranks across all of the tests in the current experiments. Additionally, further comparisons showed a strong and positive relationship between resolution (i.e., *d*=) and the difference in percentile ranking between the corrected-feedback score (optimal- bias strategy) versus the corrected-forced score (forced-report strategy): Examinees with poorer-than-average resolution achieved higher rankings at forced report whereas examinees with better-than-average resolution achieved higher rankings at optimal bias (see Figure 3). At this point a number of questions present themselves. What are the implications of these results for the validity of formula scored tests? Which measure is the best one to measure performance? If a subset of students learns to perform at optimal bias and score above other students, is that advantage unfair because it is based on a test-writing strategy? Although students tend to be overly conservative when writing formula scored tests, is the ranking that is derived from these scores still an accurate reflection of their true abilities?

Answers to the above questions are not trivial and, in our view, they depend on how these scores relate to actual academic performance in school and university, which standardized tests are presumably meant to be predicting. For example, we have demonstrated that the corrected-feedback score rank is influenced not just by retrieval (*f* in the SDT model), but also resolution. The influence of resolution on the test scores can be eliminated if examinees answer all the questions, regardless of whether the test is formula scored or not. Clearly this is a good idea from examinees’ point of view if their resolution is poor. However, whether forced report makes the test more valid depends on whether resolution, as measured by the application of SDT to formula- scored tests, predicts some aspect of academic performance.

A growing literature strongly suggests that metacognitive processes underpin important learning decisions, such as how to allocate study time (e.g., Metcalfe & Kornell, 2005) and whether to reread passages of text (e.g., Rawson & Kintsch, 2005). Thus, it is conceivable that the metacognitive processes that formula- scored tests are tapping overlap with those that underpin academic success. To the extent that this overlap exists, formula-scored tests may be more valid if examinees do not adopt an answer-all strategy. However, in our view, nonoptimal bias is unlikely to predict much of interest in school and university, so if students are to be persuaded to omit answers so that the resolution component influences the score (and thus potentially increase the validity of the test), they should be trained to perform at optimal bias rather than at their natural, overly conservative level of bias.

The Efficacy and Long-Term Effects of Feedback

Although optimal-bias feedback improved participants’ performance in Experiment 1 and 2 (i.e., corrected-feedback score > corrected score), the lasting benefits of that feedback were limited because participants showed no significant reduction in their actual bias from the first to the last test. However, the feedback manipulation in Experiment 3 *did* have lasting effects on performance: Providing feedback to participants in the educated group about optimal bias from Test 1 to guide their performance on Test 2 caused them to set a more liberal, closer-to-optimal criterion on the third test. This result is promising because it suggests that optimal-bias feedback, even if that feedback is based on a different test from the one currently being written, carries over to new tests.

The basic paradigm of Experiment 3 is exactly the scenario that would be required if training programs are to be developed to assist examinees to better prepare for upcoming formula-scored examinations such as the SAT Reasoning Test. Although there are different companies (e.g., Kaplan Guide, Princeton Review) de- signed to coach students who are preparing for large-scale tests (some of which are formula-scored), to our knowledge none focus specifically on coaching omission behavior. The only advice offered is that if examinees do not know the answer to a question, but they do know that one or more alternatives are wrong, they should answer the question. Our results have clearly shown that, although this particular piece of advice may be appropriate (i.e., scores are likely to improve under such circumstances), there is much more to the problem than is offered by this advice. That is, we have shown that the optimal omission rate is a complex construct and that it is related to prior knowledge, the penalty on the test, and resolution. Fortunately, our methodology provides the tools to estimate the optimal omission rate in a fairly straightforward manner.

Because of the number of factors that affect the optimal omission rate, examinees wanting to truly benefit from feedback may have to learn the lessons of feedback at a nondeclarative or nonanalytic level, in much the same way that other forms of complex knowledge are learned. For example, a large body of research has shown that participants exposed to consonant strings that conform to a complex rule structure (e.g., finite-state gram- mar) are later able to discriminate between test strings that also conform to that grammar and ones that do not (e.g., Reber, 1993; Higham, 1997). However, these participants are unable to verbally describe the rules that underlie such decisions (i.e., they lack declarative knowledge). Similarly, Durso and Shore (1991) showed that participants possess tacit knowledge about word meanings; for example, even participants who denied that a legal English word was part of the English language still were able to discriminate between correct versus incorrect applications of the word. These studies suggest that efforts to teach complex knowledge (such as optimal criterion setting) that are based on the provision of declarative knowledge may be in vain. Instead, ample practice and providing examinees with the opportunity to develop a backlog of instances may be the more appropriate way to go.

Another potential problem with providing declarative knowledge to participants and expecting them to respond to it has to do with strategy implementation. For example, Hertzog, Price and Dunlosky (2008; see also Dunlosky & Hertzog, 2000, 2001) had participants learn two lists of paired associated, one with a normatively effective learning strategy (interactive imagery) and the other with a normatively ineffective strategy (rote repetition). After learning, participants demonstrated that they were sensitive to the effectiveness of the different strategies as reflected in their “strategy effectiveness” ratings. However, participants seemed un- willing or unable to implement this learning with item level judgments-of-learning. Similarly, participants in our studies may possess declarative knowledge that their responding is overly conservative, but when it comes to deciding whether to report or withhold a given answer, this knowledge is not applied. Again, it may be that to overcome problems in implementing strategies at the item level, extensive practice is needed.

Although examinees may only learn optimal criterion setting with ample practice, our research points to other factors that may facilitate this learning. For example, in Experiments 1 and 2 our participants did not carry over the feedback they received on earlier tests to later tests. However, as noted above, in Experiment 3 the participants who received feedback on Test 2 did significantly reduce their omission rate on the final test, whereas participants in the control group who did not receive feedback showed no such reduction. What accounts for the different results between experiments? One possibility is that participants were made aware in Experiment 3 that the feedback that they were receiving on Test 2 was derived from Test 1. Hence, it was highlighted for them that their overly conservative responding tendencies were a general phenomenon and not specific to the current test. However, there were other differences between Experiments 1 and 2 on the one hand and Experiment 3 on the other, so this conclusion should be treated with caution. Nonetheless, presenting feedback in a "generalist" framework may be important in persuading participants to heed any feedback they are given and to transfer their learning to new tests.

A final obstacle that may need to be overcome in training students to perform at optimal bias with tests that have high accuracy (*f*) is the impact it has on the FAR. In both Experiment 1 and 2, the increase to the FAR after responding to the feedback manipulation was greater than the increase to the HR because *f* was greater than 50%. In other words, of the few errors that participants made, a large proportion of them would need to be reported to achieve the maximum-corrected score. To the extent that students are specifically concerned about their FAR and try to minimize the reporting of errors (however few or many that might be), then persuading them to respond more liberally may be difficult. Mul- tiple feedback sessions in which overall high accuracy is made salient may go some way to overcoming this obstacle. The goal of such sessions would be to teach students that they can *afford* to have a high FAR if their accuracy is high, and that a somewhat more liberal response strategy than the one they naturally adopt likely would increase their corrected score. In other words, appropriate training may encourage students to incorporate accuracy into their reporting decisions rather than these decisions being solely based on minimizing the FAR, as appears to have occurred in the current experiments.

Conclusions

The data from the present set of experiments support the claim that criterion setting is an important contributing factor to overall performance on formula-scored tests. Further, because the strategic regulation of accuracy is an essential skill (i.e., in terms of maximizing the corrected score) and, as indicated by the results of the current experiments, is responsive to certain types of feedback, further research on the topic is required. Indeed, if formula scoring continues to be applied both to large-scale exams (e.g., SAT Reasoning Test) and to course exams taken at university, then it would be remiss not to explore further the role of criterion setting and resolution on the corrected score. Such research not only will enlighten researchers and educators as to what the corrected score informs us regarding ability (e.g., separating out monitoring skills from learned facts), but also will provide a means of helping examinees understand their own aptitude and how best to perform under such testing conditions.

References

Bisseret, A. (1981). Application of signal detection theory to decision making in supervisory control. *Ergonomics, 24,* 81–94. doi:10.1080/ 00140138108924833

Bliss, L. B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. *Journal of Educational Measurement, 17,* 147–152. doi: 10.1111/j.1745-3984.1980.tb00823.x

Bruno, D., Higham, P. A., & Perfect, T. J. (2009). Global subjective memorability and the strength-based mirror effect in recognition mem- ory. *Memory & Cognition, 37,* 807– 818. doi:10.3758/MC.37.6.807

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement, 30,* 277–291. doi:10.1111/j.1745-3984.1993.tb00427.x

Cañal-Bruland, R., & Schmidt, M. (2009). Response bias in judging deceptive movements. *Acta Psychologica, 130,* 235–240. doi:10.1016/j

.actpsy.2008.12.009

Cross, L. H., & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. *Journal of Educational Measurement, 14,* 313–321. doi:10.1111/j.1745-3984.1977

.tb00047.x

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effective- ness from task experience. *Psychology and Aging, 15,* 462– 474. doi: 10.1037/0882-7974.15.3.462

Dunlosky, J., & Hertzog, C. (2001). Measuring strategy production during associative learning: The relative utility of concurrent versus retrospective reports. *Memory & Cognition, 29,* 247–253. doi:10.3758/BF03194918

Durso, F. T., & Shore, W. J. (1991). Partial knowledge of word meanings. *Journal of Experimental Psychology: General, 120,* 190 –202. doi: 10.1037/0096-3445.120.2.190

Ebel, R. L. (1968). Blind guessing on objective achievement tests. *Journal of Educational Measurement, 5,* 321–325. doi:10.1111/j.1745-3984

.1968.tb00646.x

Goldsmith, G., & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. Benjamin & B. Ross (Eds.), *The psychology of learning and motivation. Vol. 48*: *Memory use as skilled cognition* (pp. 307–324). San Diego, CA: Elsevier.

Goldsmith, M. (2011). Quantity-accuracy profiles or type-2 signal detec- tion measures? Similar methods toward a common goal. In P. A. Higham & J. P. Leboe (Eds.), *Constructions of remembering and metacognition: Essays in honour of Bruce Whittlesea* (pp. 128 –136). Basingstoke, UK: Palgrave-MacMillan.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732–769. doi:10.2307/2281536

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Hertzog, C., Price, J., & Dunlosky, J. (2008). How is knowledge generated about memory encoding strategy effectiveness? *Learning and Individual Differences, 18,* 430 – 445. doi:10.1016/j.lindif.2007.12.002

Higham, P. A. (1997). Dissociations of grammaticality and specific simi- larity effects in artificial grammar learning. *Journal of Experimental Psychology: Learning Memory & Cognition, 23,* 1029 –1045. doi: 10.1037/0278-7393.23.4.1029

Higham, P. A. (2007). No special K! A signal-detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psy- chology: General, 136,* 1–22. doi:10.1037/0096-3445.136.1.1

Higham, P. A. (2011). Accuracy discrimination and type-2 signal detection theory: Clarifications, extensions, and an analysis of bias. In P. A. Higham & J. P. Leboe (Eds.), *Constructions of remembering and meta- cognition: Essays in honour of Bruce Whittlesea* (pp. 109 –127). Basingstoke, UK: Palgrave-MacMillan. doi:10.1057/9780230305281

Higham, P. A., & Arnold, M. M. (2007a). How many questions should I answer? Using bias profiles to estimate optimal bias and maximum score on formula-scored tests. *European Journal of Cognitive Psychology, 19,* 718 –742. doi:10.1080/09541440701326121

Higham, P. A., & Arnold, M. M. (2007b). Beyond reliability and validity: The role of metacognition in psychological testing. In R. A. DeGregorio (Ed.), *New developments in psychological testing* (pp. 139 –162). Haup- pauge, NY: Nova Science.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no recognition and two-alternative-forced-choice recog- nition memory. *Journal of Experimental Psychology: General, 138,* 291–306. doi:10.1037/a0015525

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490 –517. doi:10.1037/0033-295X.103.3.490

Lueddeke, S. E., & Higham, P. A. (2011). Expertise and gambling: Using type-2 signal detection theory to investigate differences between regular gamblers and non-gamblers. *The Quarterly Journal of Experimental Psychology, 64,* 1850 –1871. doi:10.1080/17470218.2011.584631

Luna, K., Higham, P. A., & Martin-Luengo, B. (2011). The regulation of memory accuracy with multiple answers: The plurality option. *Journal of Experimental Psychology: Applied, 17,* 148 – 158. doi:10.1037/ a0023276

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Maddox, W. T. (2002). Toward a unified theory of decision criterion learning in perceptual categorization. *Journal of the Experimental Anal- ysis of Behavior, 78,* 567–595. doi:10.1901/jeab.2002.78-567

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal gamma coefficient measure of association: Implica- tions for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 509 –527. doi: 10.1037/a0014876

Meissner, C. A., & Kassin, S. M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. *Law and Human Behavior, 26,* 469 – 480. doi:10.1023/A:1020278620751

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study allocation. *Journal of Memory and Language, 52,* 463– 477. doi: 10.1016/j.jml.2004.12.001

Muijtjens, A. M. M., van Mameren, H., Hoogenboom, R. J. I., Evers,

J. L. H., & van der Vleuten, C. P. M. (1999). The effect of a “don’t know” option on test scores: Number-right and formula scoring com- pared. *Medical Education, 33,* 267–275. http://dx.doi.org/10.1046/j

.1365-2923.1999.00292.x

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109 –133. doi:10.1037/0033-2909.95.1.109

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning & Verbal Behavior, 19,* 338 –368. doi:10.1016/S0022-5371(80)90266-2

Parasuraman, R. (1985). Detection and identification of abnormalities in chest x-rays: Effects of reader skill, disease prevalence, and reporting standards. In R. E. Eberts & C. G. Eberts (Eds.), *Trends in ergonomics/human factors II* (pp. 59 – 66). Amsterdam, The Netherlands: North-Holland.

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97,* 70 – 80. doi:10.1037/ 0022-0663.97.1.70

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. Oxford, UK: Oxford University Press.

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type 1 error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics, 70,* 389 – 401. doi:10.3758/PP.70.2.389

Sax, G., & Collet, L. (1968). The effects of differing instructions and guessing formulas on reliability and validity. *Educational and Psychological Measure- ment, 28,* 1127–1136. doi:10.1177/001316446802800411

Sherriffs, A. C., & Boomer, D. S. (1954). Who is penalized by the penalty for guessing? *Journal of Educational Psychology, 45,* 81–90. doi: 10.1037/h0053756

Slakter, M. J. (1968a). The penalty for not guessing. *Journal of Educa- tional Measurement, 5,* 141–144. doi:10.1111/j.1745-3984.1968

.tb00616.x

Slakter, M. J. (1968b). The effect of guessing strategy on objective test scores. *Journal of Educational Measurement, 5,* 217–222. doi:10.1111/ j.1745-3984.1968.tb00629.x

Stretch, V., & Wixted, J. T. (1998). On the difference between strength- based and frequency-based mirror effects in recognition memory. *Jour- nal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1379 –1396. doi:10.1037/0278-7393.24.6.1379

Thurstone, L. L. (1919). A scoring method for mental tests. *Psychological Bulletin, 16,* 235–240. doi:10.1037/h0069898

Appendix

Equations to Compute Bias Profiles

Below is a general description of and the equations needed to generate a bias profile. The first step involves expressing the corrected score as a function of the hit rate (HR), the false alarm rate (FAR), the uncorrected raw score (f), and the penalty for errors (p),

$$\text{Corrected score} = \text{HR} \cdot \text{f} - \text{p} \cdot \text{FAR} \cdot (1 - \text{f}) \tag{1}$$

Now consider the contingency table shown in Table 1. The a, b, c, and *d* cells in the table represent frequencies, but it is possible to calculate the probabilities of P(a), P(b), P(c), P(d) as follows,

$$\text{P(a)} = \text{a}/(\text{a} + \text{b} + \text{c} + \text{d}) = \text{HR} \cdot \text{f} \tag{2}$$

$$\text{P(b)} = \text{b}/(\text{a} + \text{b} + \text{c} + \text{d}) = \text{FAR} \cdot (1 - \text{f}) \tag{3}$$

$$\text{P(c)} = \text{c}/(\text{a} + \text{b} + \text{c} + \text{d}) = \text{f} - \text{P(a)} \tag{4}$$

$$\text{P(d)} = \text{d}/(\text{a} + \text{b} + \text{c} + \text{d}) = (1 - \text{f}) - \text{P(b)} \tag{5}$$

Now it is possible to express the probability of guesses as function of H, FA, and f.

$$\text{P(guess)} = \text{P(c)} + \text{P(d)} \tag{6}$$

$$\text{P(guess)} = (\text{f} - \text{HR} \cdot \text{f}) + [(1 - \text{f}) - \text{FAR} \cdot (1 - \text{f})] \tag{7}$$

$$\text{P(guess)} = 1 - \text{HR} \cdot \text{f} - \text{FAR} \cdot (1 - \text{f}) \tag{8}$$

However, because a simple equal-variance Gaussian SDT model can account for the data (see Higham, 2007, Experiment 2), we can express HR as a function of FAR as,

$$\text{HR} = \Phi(\text{zFAR} + d') \tag{9}$$

where Φ(x) refers to the probability under the standard normal distribution associated with a z score equal to x, and zFAR refers to the z score corresponding to the FA rate. This term can then be substituted for HR in Equation 1 to produce,

$$\text{Corrected score} = \Phi(\text{zFAR} + d') \cdot \text{f} - \text{p} \cdot \text{FAR} \cdot (1 - \text{f}) \tag{10}$$

The same substitution can be made for the P(guess) in Equation 8 to yield,

$$\text{P(guess)} = 1 - \Phi(\text{zFAR} + d') \cdot \text{f} - \text{FAR} \cdot (1 - \text{f}) \tag{11}$$

Thus, both the corrected score and P(guess) are now expressed as a function of the FA rate, *d*= (monitoring), the uncorrected, raw score (f) and the penalty for errors (p). It is important to note that it is the critical substitution of Φ(zFAR + *d*=) for HR in Equations 10 and 11 that allows a prediction about optimal bias setting to be made. By fixing *d*=, f, and p, and varying FAR from 0 to 1, different values of the corrected score can be obtained using Equation 10, and different values of P(guess) can be generated using Equation 11. It is then possible to plot corrected scores against corresponding P(guess) values—the *bias profile*.

Received October 16, 2012
Revision received September 6, 2013
Accepted September 10, 2013 ■