Abstract Two experiments investigated the role of metacognition in changing answers to multiple-choice, general-knowledge questions. Both experiments revealed qualitatively different errors produced by speeded responding versus confusability amongst the alternatives; revision completely corrected the former, but had no effect on the latter. Experiment 2 also demonstrated that a pretest, designed to make participants’ actual experience with answer changing either positive or negative, affected the tendency to correct errors. However, this effect was not apparent in the proportion of correct responses; it was only discovered when the metacognitive component to answer changing was isolated with a Type 2 signal-detection measure of discrimination. Overall, the results suggest that future research on answer changing should more closely consider the metacognitive factors underlying answer changing, using Type 2 signal-detection theory to isolate these aspects of performance.

Many academics follow the dictum “publish or perish,” and spend their time tweaking parameters in well-established paradigms to gather data for the next paper. Lee, instead, follows the dictum “publish only if you are convinced that you have discovered something truly new and interesting about the world.” Quite rightly, many of us should be nervous applying this philosophy to our own careers; such an approach will only work for true intellects – scientists with unique philosophy and vision. But for the special few with such vision, the result is science that is long lasting, and careers that are distinguished. From his block letter F research (Brooks, 1967), to the beginnings of instance theory (Brooks, 1978), to his more recent work in medical diagnosis (e.g., Brooks, LeBlanc, & Norman, 2000), it is clear that Lee is one of those special few.

Those of us who were lucky enough to receive Lee’s mentorship learned many lessons. One such lesson was to stay open-minded, reading more than the most immediate literature on a research topic, and trying to think beyond the current paradigm used to investigate a phenomenon. Such a strategy can sometimes lead to new research as cross talk between previously independent literatures is developed. Developing such cross talk between the recognition memory domain, and the implicit learning domain, motivated the research reported in Higham and Brooks (1997). The current research focuses on two other literatures where cross talk could be fruitful: psychological testing and the newer field of metacognition. Although substantial research exists on metacognition in educational settings, this research has been predominantly focused on how various subjective ratings (e.g., judgments of confidence, comprehension, or learning) relate to objective test performance (e.g., Maki & McGuire, 2002; Pressley, Ghatala, Woloshyn, & Pirie, 1990). Very little to no research has considered the metacognitive issues involved in actual performance on the objective tests themselves.

In this paper, we will report two experiments that examine an old issue in the testing literature – the effect of changing answers on multiple-choice tests (e.g., see Benjamin, Cavell, & Shallenberger, 1984; Crawford, 1928; Crocker & Benson, 1980; Foote & Belinki, 1972; Geiger, 1996; Harvil & Davis, 1997; McMorris, Demers, & Schwartz, 1987; Mueller & Wasser, 1977; Vispoel, 1998, 2000). Overall, this research has revealed a number of consistent findings that include: 1) a small proportion of answers on a given test are typically changed, 2) the changes usually increase the overall test score, indicating that more answers are changed from wrong to right, than from right to wrong, and, 3) most examinees change a few answers. This last finding is particularly interesting because surveys have indicated that students tend toward the belief that it is preferable to stay with first hunches, and not to make changes to their answers. Indeed, Mathews (1929) reported that 86% of his sample of students believed that changing answers lowered the overall test score. This belief may partially stem from the fact that educators often pass this information on to their students (e.g., see Foote & Belinki, 1972).
In our view, however, there is no straightforward solution to the problem of whether or not to change answers because the interplay of at least three variables is involved: 1) the number of initial answers that are correct, 2) the number of initial answers that are changed, and 3) students’ ability to discriminate between initial answers that are correct, and those that are not. Unfortunately, however, the literature on answer changing has never taken all three of these factors into account when considering students’ best revision strategy, and to our knowledge, none of it directly addresses the metacognitive or discrimination processes that may be involved.

In the experiments that we report here, we introduce a methodology that directly addresses metacognition in answer changing. In each experiment, participants were presented with a series of four-alternative, multiple-choice, general knowledge questions one at a time at a computer workstation (Phase 1). After completing the test, each question was presented again, along with the provided answer, and participants were given the opportunity to change their initial answer (Phase 2). The two experiments differed in that a paper-and-pencil pretest was used in Experiment 2, but not in Experiment 1. There were two versions of the pretest, intended to vary participants’ actual experience with changing answers before starting the main test. One version contained deceptive questions, whereas the other contained non-deceptive questions. Participants were required to change a large portion (33%) of their initial pretest answers to both question sets. We hypothesized that being required to change one-third of their answers to non-deceptive questions would result in a decrease to participants’ scores, because one-third was more than the optimal proportion (i.e., several correct to incorrect changes would occur). Conversely, being required to change one-third of their answers to deceptive questions should increase participants’ scores because several first hunches may have been wrong.

What particularly marks our research as unique, however, is the application of Type 2 signal-detection theory (SDT; e.g., Clarke, Birdsall, & Tanner, 1959; Galvin, Podd, Drga, & Whitmore, 2003; Higham, 2002; Higham & Tam, in press; Vokey & Higham, 2005), which allowed for the derivation of separate quantitative estimates of the three components of answer changing listed above. Type 2 SDT differs from standard Type 1 SDT in that participants are required to discriminate between their own correct and incorrect answers, rather than between signal-plus-noise and noise-only trials. Consequently, the discrimination index in Type 2 SDT is a measure of metacognitive monitoring.

Table 1 shows how we applied Type 2 SDT in a paradigm involving answer changing. The left and right columns represent the frequency of initially correct and incorrect answers, respectively, whereas the top and bottom rows represent the number of unchanged and changed answers, respectively. The combination of these variables yields four cell frequencies: A = unchanged, initially-correct items; B = unchanged, initially-incorrect items; C = changed, initially-correct items; and D = changed, initially-incorrect items. Because changing an initially-incorrect answer can result in either a correct or incorrect second response on any test involving more than two options, e is also included in Table 1. This value, which appears in Table 1 in lowercase to reflect the fact that it is a subset of D, represents the number of initially incorrect items that are correct after a change. These five values (i.e., A to e) allow several measures of performance to be determined. The first two measures are standard in the literature: 1) proportion correct before change = (A + C) / (A + B + C + D), and 2) proportion correct after change = (A + e) / (A + B + C + D). However, it is also possible to derive Type 2 SDT hit (A/[A + C]) and false alarm (B/[B + D]) rates, which represent the proportion of initially-correct responses that were unchanged, and the proportion of initially-incorrect responses that were unchanged, respectively. Using these rates, measures of discrimination (A; Grier, 1971) and change bias (B/D; Donaldson, 1992) were calculated. These latter two measures are unique in this context, and reflect participants’ ability to monitor the accuracy of their initial answers, and their tendency to change answers, respectively.
Participants

There were 24 participants in Experiment 1, and 48 in Experiment 2. In Experiment 2, 25 participants were assigned to the negative pretest group, and 23 to the positive pretest group.

Design and Materials

The main test in both Experiment 1 and Experiment 2 consisted of 110, four-alternative, multiple-choice, general-knowledge questions adapted from the Nelson and Narens (1980) norms. Ten of these questions were for practise and the remaining 100 were critical questions used in the analyses. For both experiments, two within-subject independent variables were manipulated, each with two levels: item difficulty and answer deadline. Item difficulty was manipulated by varying the alternatives provided with each question. For easy questions, the correct answer was included with three incorrect alternatives not easily confused with the correct answer (e.g., “Of which country is Buenos Aires the capital? [a] Canada [b] Mexico [c] Argentina [d] Cuba”), whereas for difficult questions, the confusability amongst the alternatives was greater (e.g., “[a] Columbia, [b] Brazil [c] Argentina [d] Chile”). Assignment of correct answers to the “a,” “b,” “c,” and “d” alternatives was balanced across questions. Answer deadline was manipulated by providing participants with either 5 s or 10 s to provide an answer to each question. Assignment of questions to each of these four experimental conditions was counterbalanced across participants using four test formats, with 25 critical questions assigned to each of the four cells of the design. A different random order of presentation of the 100 critical questions was used for each participant, but the same random order was used (within subjects) for the first attempt at the questions (Phase 1), and the second, revision phase (Phase 2). The 10 practise questions were presented at the beginning of Phase 1, but not Phase 2, with 5 questions each at the 5 s and 10 s deadlines.

In addition to the two within-subject variables, pretest type (positive vs. negative) was manipulated between subjects in Experiment 2. This pretest consisted of 30, four-alternative, general-knowledge, multiple-choice questions administered with paper and pencil. For the positive pretest group, the questions were designed to be deceptive, such that the most obvious answer was not actually correct. For example, for the question, “What is the capital of Australia? [a] Sydney [b] Canberra [c] Adelaide [d] Melbourne,” the correct answer is “Canberra,” although many people would choose “Sydney.” For the negative pretest group, the questions were of medium-level difficulty and non-deceptive.

Procedure

Participants in Experiment 1 started with the main computer-administered test. They were informed that they would be presented with general-knowledge questions, each with four alternatives, one at a time, on the computer. They were to choose one of the four alternative answers by pressing one of the keys labelled “a,” “b,” “c,” or “d” (which actually corresponded to the “a,” “s,” “d,” and “f” keys, respectively, on the computer keyboard). They were told that they were to provide answers within a time limit, and that a clock which counted down in seconds from either five or ten would appear on the screen along with the question. Failure to respond in the time allotted, or pressing a key not assigned to one of the four response alternatives, resulted in feedback that an error had been made, and data from these trials were not analyzed. Participants were also informed of the following point system, which was implemented to encourage question answering: +1 = correct; 0 = incorrect; -4 = no answer. No feedback on points was provided during the test.

Participants began with the 10 practise questions, and then answered the 100 critical questions. Participants were not informed that the initial trials were practise, and nothing demarcated the practise and critical testing phases. After completing all 110 questions, participants were given further instructions, and then began Phase 2. During Phase 2, participants were shown the same 100 critical questions again, one at a time, with their chosen answer indicated. They were required to type in a response to each question again, changing their original answer if desired. Unlike Phase 1, no deadline was imposed, and there was no mention of a point system. Participants were also required to rate their confidence (1 to 6) after answering each question in both Phase 1 and 2. However, these data were not analyzed.

Participants in Experiment 2 started with the paper and pencil pretest, and then proceeded to the main test, which was exactly the same as in Experiment 1. After completing the pretest answer sheet in blue ink, participants were required to engage in a distracter task (word puzzles) while the experimenter marked their test. The answer sheets were then returned to participants without feedback, and they were instructed to choose 10 of their 30 answers (33%) and change their answers to them, using a red pen, so that changes could be easily identified. After the changes were complete, the experimenter marked the pretest again, while participants continued with the distracter task. Participants were then told their total correct score before and after changing answers.
Results

Alpha level of .05 was adopted for all comparisons. The proportion of correct responses on the main test in Experiments 1 and 2, as a function of question difficulty, deadline, and test phase, is shown in Figure 1. The data from Experiment 1 were submitted to a 2 (Difficult/Easy) x 2 (10 s/5 s) x 2 (Phase 1/Phase 2) within-subjects analysis of variance (ANOVA), which revealed significant main effects of question difficulty, $F(1,23) = 64.19$, $MSE = .017$ (difficult: .68; easy: .83), test phase, $F(1,23) = 43.13$, $MSE = .009$ (Phase 1: .71; Phase 2: .80), and a deadline by phase interaction, $F(1,23) = 10.18$, $MSE = .005$. The interaction was caused by the fact that the opportunity to revise answers had a larger effect on items initially answered at 5-s deadline than at 10-s deadline. The data from Experiment 2 were analyzed with a similar ANOVA, only with the between-subject factor pretest type (Positive/Negative) added. This ANOVA, too, revealed main effects of question difficulty, $F(1,46) = 152.94$, $MSE = .015$ (difficult: .65; easy: .81), test phase, $F(1,46) = 45.95$, $MSE = .010$ (Phase 1: .70; Phase 2: .77), and a deadline by phase interaction, $F(1,46) = 8.53$, $MSE = .006$. The interaction reflected the same data pattern as in Experiment 1. The main effect of deadline was also significant, $F(1,46) = 4.29$, $MSE = .028$ (10 s: .75; 5 s: .71). Importantly, the analysis revealed no effects of pretest, which is why the data for Experiment 2 are presented in Figure 1 collapsed across this factor.

A 2 (Positive/Negative) x 2 (Before Change/After Change) mixed ANOVA on the number of correct responses on the pretest revealed a significant interaction, $F(1,46) = 7.47$, $MSE = 3.09$. As predicted, in the positive pretest group, answer changing increased the mean number of correct answers on the pretest from 16.30 (54%) to 16.82 (56%), whereas it decreased it in the negative pretest group from 18.16 (61%) to 16.72 (56%).

Before calculating the Type 2 SDT measures of monitoring ($A'$) and change bias ($B''D$) on the main test, the hit and false alarm rates were adjusted to avoid undefined values (i.e., hit rate = $[A + 0.5] / [A + C + 1]$;...
false alarm rate = \[ \frac{B + 0.5}{B + D + 1} \]; Snodgrass & Corwin, 1988). These data are shown in Table 2. A 2 (Difficult/Easy) x 2 (10 s/5 s) within-subjects ANOVA on monitoring in Experiment 1 revealed a main effect of deadline, \( F(1,23) = 7.17, \) \( MSE = .004 \) (10 s: .82; 5 s: .85), and a main effect of question difficulty, \( F(1,23) = 21.92, \) \( MSE = .005 \) (easy: .87; difficult: .80). Additional analyses determined the effect, if any, of pretest type on monitoring in Experiment 2. Although the analysis of proportion-correct data revealed no effect of pretest, a 2 (Difficult/Easy) x 2 (Positive/Negative) mixed ANOVA on monitoring revealed a main effect of question difficulty, \( F(1,46) = 24.41, \) \( MSE = .009 \) (easy: .85; difficult: .76), and a pretest by difficulty interaction, \( F(1,46) = 5.09, \) \( MSE = .010 \). As can been seen in Table 2, a negative pretest lowered monitoring relative to a positive one, but only on easy questions. Analogous ANOVAs on change bias revealed a main effect of deadline in Experiment 1, \( F(1,23) = 13.18, \) \( MSE = .010 \) (10 s: -.82; 5 s: -.65), indicating that participants changed answers to more questions initially answered with the 5-s deadline than the 10-s deadline. No effects on change bias were significant in Experiment 2, although the largest F value, \( F(1,46) = 2.40, \) \( MSE = .055, p > .15 \), was also associated with the main effect of deadline.

Discussion

The current experiments were conducted to determine the role of metacognition in changing answers on multiple-choice tests. Consistent with previous research, providing participants with the opportunity to revise their answers increased their overall test score. However, the manipulations of question difficulty and answer deadline showed that not all errors were correctable; revision in Phase 2 allowed participants to completely eliminate any disadvantage caused by the short versus long answer deadline in Phase 1. Conversely, questions with confusable alternatives remained harder than questions with less confusable alternatives even after participants were given the opportunity to change their answers (Figure 1). This finding may help to explain the paradox that participants simultaneously hold the opinion that correcting answers is unbeneficial, but generally do change a few answers when given the opportunity to do so, and generally improve their score in the process. Errors like those produced by an answering deadline (e.g., typographical errors, question misinterpretation or misreading) are likely to be monitored well, are likely to be corrected, and their correction should function to improve the test score. However, they are not likely to receive much attention. In contrast, students are likely to “sweat over” questions with confusable alternatives and check them for accuracy when receiving feedback on their test performance. But because there is little improvement to changing answers to these questions, the possible result is that students form the erroneous belief that answer changing never produces benefits.

Examining only the proportion of correct answers in Experiment 2 did not reveal any effect of pretest. Indeed, had performance only been analyzed in terms of whether or not participants produced correct responses, which is typical in the testing literature, we would have concluded that pretest had no effect. However, proportion correct can be a confounded measure of performance. It may be influenced not just by how much participants know, but also by their monitoring ability and change bias. By separating out the monitoring component, using a Type 2 index of discrimination, the effect of pretest was revealed; after receiving negative pretest feedback, participants did not monitor the errors that they made to easy questions as well as participants given positive pretest feedback (Table 2). The precise explanation for this result will have to wait for follow-up research, but it is worth speculating that the experience of negative feedback caused participants to recruit negative prior instances of answer changing, or of instances of being told by instructors to stay with “first hunches.” Recruitment of such negative instances may have led participants in the negative pretest group to leave more incorrect answers unchanged, compared to the positive feedback group. Regardless of the exact reason for the monitoring decrease, however, it is clear that traditional methods are not suitable for investigating metacognition in answer changing, and that something like Type 2 SDT will be necessary in the long run.

Philip A. Higham and Catherine Gerrard, School of Psychology, University of Southampton. Portions of this research were presented at the 13th annual meeting of the Canadian Society for Brain, Behaviour and Cognitive Science, June, 2003 at McMaster University, Hamilton, Ontario, Canada. Thanks to John Vokey, Scott Allen, and Sam Hannah for comments on an earlier draft of this paper. PH was Lee Brooks’ graduate student from 1987 to 1992. He was awarded his PhD in 1993.

Correspondence concerning this article should be addressed to Philip A. Higham, School of Psychology, University of Southampton, Highfield, Southampton, UK SO17 1BJ (E-mail: higham@soton.ac.uk).

References


Deux expériences ont servi à faire enquête sur le rôle de la métacognition dans le changement des réponses données à des questions à choix multiple faisant appel aux connaissances générales. Dans le cadre de la première expérience, les participants ont d'abord répondu à 100 questions à quatre réponses possibles. Deux variables indépendantes étaient déterminées pour chaque sujet, c'est-à-dire la difficulté des questions et le délai de réponse. La difficulté des questions était réglée par l'inclusion de réponses incorrectes susceptibles ou non d'être confondues avec la bonne. En ce qui concerne le délai de réponse, le répondant avait soit cinq soit dix secondes pour répondre à chaque question. Après avoir répondu à toutes les questions à la première étape, les participants avaient la possibilité, à la deuxième étape, de changer leurs réponses à chacune des questions et disposaient de tout le temps voulu pour le faire. À cette fin, chaque question et les différentes réponses possibles leur étaient présentées à nouveau, de même que leur réponse initiale. Soit que les participants ne modifiaient pas la première réponse donnée (même réponse) soit qu'ils
choisissaient une nouvelle réponse (réponse changée). Les résultats ont montré que les erreurs commises à cause du délai de réponse étaient tout à fait corrigeables et que l'avantage de ceux qui bénéficiaient d'un délai de réponse de dix secondes plutôt que cinq à la première étape était complètement éliminé à la suite de la révision des réponses à la deuxième. Par contre, les erreurs causées par la difficulté des questions n'étaient pas corrigeables, si bien que l'avantage procuré par les questions faciles en comparaison des questions difficiles à la première étape est demeuré entier à la suite de la révision des réponses à la deuxième. Ces résultats nous apprennent que la compression du temps de réponse et le potentiel de confusion entre les réponses possibles ont donné lieu à des erreurs différentes du point de vue qualitatif. Lors de la deuxième expérience, les participants ont passé un test préliminaire avant de procéder au test (et de bénéficier de la possibilité de revoir les réponses) prévu par la première expérience. L'objet du test préliminaire était de conférer aux participants une expérience réelle positive ou négative du changement de réponses. À cette fin, des questions trompeuses ou non trompeuses étaient intégrées au test préliminaire, et les participants se sont donc trouvés dans l'obligation de changer une part élevée des réponses. Par leur conception, les questions trompeuses suscitaient des réponses initiales intuitives fausses, de telle sorte que le remplacement de nombreuses réponses tendait à augmenter la note obtenue. En ce qui concerne les questions non trompeuses, le changement de nombreuses réponses tendait à diminuer la note, comme bien des bonnes réponses étaient remplacées par des réponses erronées. L'analyse de la part des bonnes réponses données au test principal donne à entendre que l'effet de l'expérience du test préliminaire était nul ; la part des bonnes réponses données au test principal était pratiquement identique à la part obtenue lors de la première expérience, qui ne prévoyait aucun test préliminaire. Par ailleurs, outre la mesure de la part des bonnes réponses, une mesure de discrimination de type 2 fondée sur la détection des signaux a servi à isoler l'élément de contrôle métacognitif qui jouait dans le changement de réponses. L'élément en question était une mesure de la tendance des participants à changer les réponses erronées et à ne pas toucher aux bonnes réponses. Les analyses de l'élément ont révélé que, contrairement au test préliminaire positif, le test négatif diminuait le résultat du contrôle des questions faciles, peut-être du fait que les participants avaient tendance à ne pas toucher aux réponses incorrectes qui leur étaient venues spontanément par intuition. Dans l'ensemble, les résultats indiquent que les recherches futures sur les changements de réponses devraient prendre en compte davantage les facteurs métacognitifs qui sous-tendent le changement et faire appel à cette fin à la théorie de détection de signaux de type 2 afin de dégager cet aspect de la performance.