ASSESSMENT CERTITUDE AS A FEEDBACK STRATEGY FOR LEARNERS' CONSTRUCTED RESPONSES

By: W. A. KEALY and A. D. RITZHAUPT

Kealy, W. A. & Ritzhaupt, A. D. (2010). Assessment certitude as a feedback strategy for learners' constructed responses. *Journal of Educational Computing Research*, 43(1), 25-45. doi:10.2190/EC.43.1.c

Abstract:

Educational researchers have rarely addressed the problem of how to provide feedback on constructed responses. All participants (N= 76) read a story and completed short-answer questions based on the text, with some receiving feedback consisting of the exact material on which the questions were based. During feedback, two groups receiving feedback also rated the certainty of their response correctness—an activity we have termed assessment certitude, Additionally, participants in one of these groups viewed their initial responses along with the feedback. All three feedback conditions showed significant gains in recall performance compared to a fourth group that received no feedback. Low ratings of assessment certitude significantly correlated with improved recall for two groups receiving feedback that did not include their original responses. Among these participants, mental reiteration of the feedback received was the most frequently used mental strategy while participants in a third group, who saw their original responses during feedback, reported using other approaches such as visualization.

Article:

Feedback has been shown to be one of the most effective instructional interventions for improving learning (Hattie & Timperley, 2007). It has also been a subject of enduring interest for educational researchers, and is a key component in behaviorist and cognitive theories of learning and related research (BangertDrowns, Kulik, Kulik, &Morgan, 1991). One consensus about feedback that has been building among researchers is that effective feedback should include *verification* about the correctness of a response as well as an *elaboration* that provides learners with relevant information for constructing a correct response (Kulhavy &Stock, 1989; Mason &Bruning, 2001; Mory, 2004; Shute, 2008).

One longstanding issue in the use of feedback is the difficulty in verifying the correctness of learner responses to criterion measures consisting of short-answer or essay questions that are difficult to score by machine (Braun, Bennett, Frye, & Soloway, 1990; Leacock & Chodorow, 2003). Despite the advances of natural language processing in the computing sciences, it is still difficult to provide automated correctness feedback on constructed-response items with high reliability. This explains, in part, why the vast majority of research studies on feedback has employed criterion measures such as multiple-choice or true- false questions that are clearly either correct or incorrect (Kulik & Kulik, 1988, Kulhavy &Stock, 1989; Mory, 2004). The tradeoff for this accuracy in assessment, however, is that learning performance can only be measured in terms of how well one *recognizes* a correct response (among a set of distracters) rather than what one *remembers* as an expression of cued or free recall. The present research was therefore informed by the practical need (Reeves, 2000) to find approaches for providing feedback to short-answer responses when there is no verification of response correctness.

Given the absence of a means for verifying correctness of response, we focused on the elaboration component of feedback, which generally contributes to better learning than verification (Bangert-Drowns et al., 1991, Mason &Bruning, 2001; Shute, 2008). Specifically, we examined the effectiveness of a form of elaboration similar to what Mason and Bruning (2001) called “topic-contingent” feedback consisting of material on which a given question is based. The strategy of simply representing target material as feedback for a related question is justified by research indicating that learners who are given the opportunity to reread the text from which an answered question was derived perform better on a retest containing the same questions than those who are either provided just the correct answer or given no feedback at all (Andre & Thieman, 1988).

Fundamental to the elaborative feedback strategy explored herein is the view that learners are themselves an important source of feedback (Bangert-Drowns et al., 1991; Butler &Winne, 1995; Hattie & Timperley, 2007). Internally- generated feedback includes estimation of performance relative to a goal, judgments on the efficacy of learning strategies and tactics used, and the feelings and beliefs associated with self-efficacy and ability (Butler &Winne, 1995). Moreover, depending on the knowledge, beliefs, metacognitive strategies, and motivations of the learner, internally-generatedfeedback and the goals, strategies, and tactics used during the learning process will vary. In a sense, the distinction between external and internal feedback is arbitrary since the former typically elicits some variety of the latter. Hence, we view feedback that presents relevant target text following a question response as more than simply a second exposure to instructional content. Rather, it is a content correction that serves as the basis for internal processing and self-regulated learning (SRL).

THE ROLE OF SELF-REGULATED LEARNING IN FEEDBACK

Self-regulation involves activation and maintenance of specific processes that are adapted to a learning task including goal-setting, use of metacognitive strategies, time management, self-evaluation, and monitoring of performance (Schunk, 2008; Zimmerman, 2002). Mory (2004) contended that SRL is the “missing link” between findings from studies on feedback and the motivational and constructivist factors in learning. Indeed, a new view of feedback has emerged that situates feedback within a SRL context (Butler &Winne, 1995; Moos & Azevedo, 2008; Mory, 2004). Two important aspects of this perspective are the recursive nature of SRL (Zimmerman, 2008a) and the importance that self-monitoring plays in this process. Feedback internally generated during monitoring of performance guides subsequent self-regulation. Hence, feedback not only performs a corrective function by providing domain information but also works to guide cognitive activities for processing this information and constructing knowledge (Butler &Winne, 1995). The latter, however, is impossible without monitoring —the critical component of SRL—in which learners continually assess their state of progress on a task and compare this to their goals. The discrepancies that arise from such comparisons during monitoring generate internal feedback that, in turn, modifies the learning strategies and tactics one uses.

For this study, we particularly focused on the self-assessment component of SRL. Self-assessment may provide evidence of increased competence, thereby having a positive impact on learner performance (Schunk, 1996; Schunk & Ertmer, 1999). Besides this affective role, self-assessment may be considered a self-regulatory practice that involves interpreting external feedback and evaluating the present state of knowledge, abilities, and cognitive strategies (Hattie & Timperley, 2007). In line with this view, our study examined the effectiveness of a self-assessment activity designed to enhance the value of elaborative feedback (i.e., target text associated with ashort-answer question) and positively influence the selection and use of learning strategies and tactics. This assessment activity, involving rating the certainty of response correctness concurrently with feedback—which we have termed *assessment certitude*—was implemented with two purposes in mind. Our primary intention was to examine a scheme for guiding performance on constructed-response tasks. A secondary purpose of the study was to seek a better understanding, as some researchers have urged (Butler & Winne, 1995; Zimmerman, 2008a), of the self-regulation and learner adaptation that occurs following external feedback.

The work of Kulhavy and Stock (1989) represents an early attempt to investigate the mental processes underlying feedback and explain differences in the effects of feedback. This inquiry was accomplished by introducing the notion of *response certitude,* defined as a metacognitive knowledge estimate of whether an instructional response is correct. Response certitude is a component in the Kulhavy-Stock model where feedback is a system input which, when compared with a cognitive referent (i.e., recall of one’s initial response to an instructional task plus the certainty of that response), results in a discrepancy value that the feedback system attempts to resolve by modifying the memory of the initial response. A discrepancy value is non-existent if learners verify the correctness of a response that was made with a high degree of certitude. By contrast, a high discrepancy value results when learners realize their answer was wrong, despite their initially high response certitude. A relatively higher discrepancy forces the system to work harder to achieve equilibrium, yielding a stronger memory trace during feedback study and better recall performance on a future test. This view, reasoned Butler and Winne (1995), is consistent with their model of feedback as self-regulation in which "monitoring (perceptions of discrepancy) influences goals students set, which affects subsequent cognitive tactics applied (time spent processing feedback), influencing performance (correcting erroneous responses)" (p. 270).

Besides its usefulness as an analytical tool for understanding the nature of feedback, we believe that the act of having people report their response certitude also performs an instructional function that strengthens the encoding of one's initial response. By requiring learners to retrieve and mindfully evaluate their responses, the memory trace for the initial response is strengthened, thereby more easily bringing the response into working memory for comparison with the feedback given. An improved ability to recall one's initial response to a test item could be especially useful in cases where feedback consists of just the original text from which the assessment item was taken. It was with this realization that we were inspired to adapt the idea of response certitude, reconceptualizing it for the present study as assessment certitude.

Kulhavy and his colleagues obtained data on response certitude by having research participants rate confidence on a 5-point Likert scale (ranging from "random guess" to "absolutely sure") immediately before they responded to a question (e.g., Kulhavy, Yekovich, &Dyer, 1976). By contrast, the *assessment certitude* incorporated in the current study was measured at the time feedback (i.e., target text on which the constructed response item was based) was presented. Incorporating an assessment certitude activity with feedback, we reasoned, should increase engagement on the mental task and promote SRL (Hattie & Timperley, 2007). Further, while Mory (2004) suggested that response certitude may indicate more than simply what learners know but also reflect their general level of motivation and self-esteem, these factors should have little or no influence on assessment certitude that involves direct comparison between one's response to a question and the correct information. This scheme of presenting elaborated feedback simultaneously with an assessment certitude rating task approximates Butler and Winne's (1995) idea of functional validity feedback that "describes the relation between the learner's estimates of achievement and his or her actual performance" (p. 252).

Two groups of participants in the experiment viewed elaborative feedback while concurrently rating the certainty of their previous responses. A third group viewed the feedback but did not make an assessment certitude rating on each item, and a fourth group received no feedback. Among those viewing feedback and completing an assessment certitude rating task, one group of participants also viewed their initial response alongside the paragraph from which the question was formed. The rationale for including this third group in the study was to determine whether being able to make a direct comparison between response and feedback might contribute to more accurate estimates of one's performance and, in turn, better self-monitoring (Butler &Winne, 1995). Alternatively, making an original response available for side-by-side comparison with target text could result in reduced mental effort that undermines monitoring.

We framed three hypotheses that served as the basis for the current experiment:

1. participants who performed assessment certitude ratings would exhibit better recall than those who either saw just the relevant target text as feedback after answering a question or viewed no feedback;
2. those seeing their original response concurrently with the target text while rating their assessment certitude would generate more accurate and lengthier responses than those who viewed only the target text while making their rating; and
3. participants viewing target text following response to a question would exhibit superior text recall than those in a group receiving no feedback.

METHOD

Research Design and Participants

In this experiment, we compared text recall of a prose passage under four different feedback conditions: a) a repeat presentation of the sentence in the experimental text that was the source of the question answered (F); b) the same type of feedback along with a concurrent rating task to assess the accuracy of the answer given (FA); c) feedback and self-assessment rating while one's original response was in full view (FAR); and d) no feedback (NF). Two testing occasions occurred, one prior to feedback and the other following feedback and after the reading of a short unrelated passage designed to clear short-term memory (STM), Hence, the study followed a pretest-posttest control group experimental design (Campbell &Stanley, 1966) in which Feedback Type was the independent variable with participants in the NF group serving as controls. The primary dependent measure was improvement in text recall performance between the two testing sessions (hereafter referred to as test 1 and test 2). This criterion measure consisted of participants' accuracy in their written responses and the length of them. Additionally, we measured the time spent studying the feedback and how certainty level associated with each answer given.

Participants were 76 undergraduate education majors at a large southern U.S. research university. These volunteers, who received extra credit for their involvement, were randomly assigned to computers yielding this distribution among the four experimental treatments: F = 19, FA = 17, FAR = 21, NF = 19.

Materials

A fictitious story titled "The Roman Town of Albano" (word count 631) was used that had a Flesch Reading Ease of 62.3 and a Flesch-Kincaid Grade level of 8.5. The story was presented on a sequence of 14 computer screens via a program that had been created using Authorware 4.0. The first screen contained a two-sentence introduction while the last presented a one-sentence summary. Most of the story appeared on the 12 intervening screens, each consisting of a three-sentence paragraph with the second and third sentences containing information from which the 24 criterion questions were generated.

Tiny arrows appeared at the bottom of every screen that were clicked to navigate through the story. A notation in the upper-left corner of each screen indicated the reader's place in the 14-page story (e.g., the third screen was annotated "pg, 3/14,"), A numerical counter in the upper-right corner showed how much time (in seconds) of the 5 minutes allocated for reading the story remained. Figure 1 illustrates the typical screen layout.

Two versions of the experimental text were created that presented the same information, but in a different order. The purpose of this was to exclude the possibility of differential performance due to the order in which the prose was presented; we expected no significant variability in recall performance as a result of order effects. Given the study's four between-subjects conditions, this arrangement yielded eight computer programs that were installed, in equal numbers, on 16 PCs in a computer lab.

Procedures

Participants were randomly assigned to a computer containing one of the four experimental treatments. After receiving an overview of what they would be doing, the researcher addressed all procedural questions and instructed them to start the program (by clicking the Tab key). The program indicated that the next step involved reading a short story of 12 paragraphs for which 5 minutes were allocated to complete the task.

FIGURE 1 IS OMITTED FROM THIS FORMATTED DOCUMENT

After reading, participants completed three two-column addition problems on their computers for the purpose of clearing working memory. The program instructed them that they would be answering 24 short-answer questions based on the text completed. Participants in the three feedback conditions (F, FA, and FAR) were informed that, following the response to each short-answer question, they would be shown the story sentence on which the question was based and that they were to use this information to judge the correctness of their answer.

Those receiving NF were only instructed to type a short answer in the space provided for each of the 24 questions (and then press Enter).

Regarding the FA and FAR groups, the computer instruction added that their accuracy judgments would be made using an onscreen slide. Participants practiced using the onscreen slide. The slide consisted of a narrow horizontal window at the bottom third of the screen with a thin vertical bar (the slider) that could be slid left and right. The window was marked off in five sections by six tick-marks below the slide bar labeled 0%, 20%, 40%, 60%, 80%, and 100%. At the far left and right ends of the slide window appeared, respectively, the words totally inaccurate and totally accurate. An instruction to "Grade the percentage of accuracy of your response" was printed above the slide window while below it a button labeled "Click to record score and view next question" was shown. Figure 2 depicts the slide that participants used to rate the accuracy of their responses. For each rating the slide could be moved, released, and then repositioned to reflect a change in judgment. Not until participants clicked the button located below the slide did the computer collect and store the rating made.

After completing the 24 constructed-response items, participants read a 276- word passage titled "The Lakes and Fishes of Albano," This passage, presented over four computer screens, appeared between the first and second cued recall tests to clear STM, thereby revealing the more durable effects of feedback on recall. Once participants finished the reading, they again completed the constructed-response test, but without any feedback or self assessment. On both test occasions, the order in which questions appeared was randomized for each participant. Following the second test, the computer prompted participants to report any mental tricks or strategies they used for recalling the story. They typically completed the experimental session in 30 minutes.

FIGURE 2 IS OMITTED FROM THIS FORMATTED DOCUMENT

RESULTS

During scoring, one point was assigned to a response capturing the gist of the correct answer and two points were awarded for correct responses that were more elaborate. The two authors independently scored identical samples of the protocols, resolving scoring differences until inter-rater agreement exceeded 90%. An alpha level of ,05 was used for all tests of significance.

Recall data from the two versions of the story were first compared with a one-way ANOVA. Results showed no significant differences in performance due to prose order used during either test 1,$F(1, 74) = 0.68$, $p = .41$,d =.13, or test 2, $F(1, 74) = 0.53$, $p = .47$,$d = .18$. Consequently, we pooled the data from the two text versions for further analysis.

Recall Performance

Participants' relative performance in recalling material from the story was determined through differences between the first and second tests along two measures: correctness of response and response productivity (i.e., the length of participants' answers).

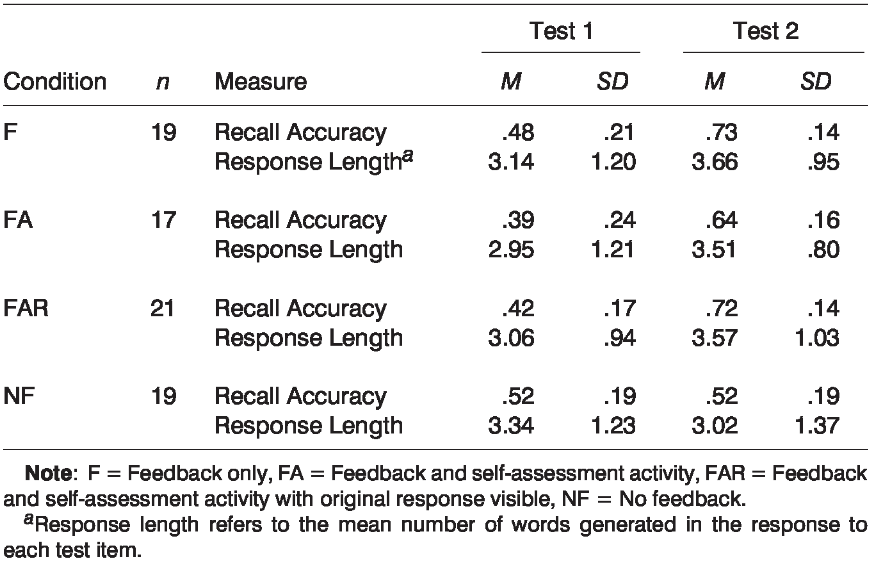
Correctness of Response

The mean percentage of correct recall by the four experimental conditions during test 1 and test 2 is presented in Table 1. Participants in the FAR group showed the greatest gain in recall score (30%) with only slightly lower increases shown by those in the two other feedback conditions (both 25%). By contrast, the NF group showed no improvement in recall accuracy during the second test.

A one-way ANOVA on recall performance during test 1 showed no significant difference in ability among the four treatment groups; $F(3, 72) = 1.54$, $p = .21$. Performance on test 1 was subtracted from that of test 2 to compute gain scores for the four groups used in subsequent analysis. Analysis of gain scores, we reasoned, was preferable over an analysis of test 2 scores because the former afforded a more intuitive grasp of the relative impact of the different treatments. Although some researchers (e.g., Gay & Airasian, 2000) question the validity of using gain scores (high performers on a pretest have less opportunity to improve), we believed such concerns were unwarranted since, in this instance, the participants were not familiar with the content of the experimental text. Further, no differences were evident between groups in their test 1 performances and a significant, $r = .62, p < ,01$, correlation existed between test 1 and 2 scores.

Results of a one-way ANOVA revealed that the four groups differed significantly, $F(3, 72) = 23.61, p < .01, \eta^2 = .50$, in improved recall performance on test 2. Bonferroni pairwise comparisons of the adjusted scores revealed that participants in all three feedback conditions significantly outperformed those not receiving feedback. To draw out distinctions among the three feedback treatments, deviation contrasts were performed on the data in which each condition, except the last category in the dataset (i.e., the NF group), was compared to the overall effect. Significant differences arose in contrasts involving the F group, $t(36) = 2.22$, $p = .03, d = .59$, and FAR group, $t(38) = 4.13, p < .01, d = .98$, whereas the contrast for the FA group was marginally significant, $t(34) =1.96$, $p = .05, d = .49$. While contrasts for the F and FA groups showed medium effects sizes, a large effect size was associated with gains by the FAR group.

Table 1. Mean Percent Correct Recall and Length of Responses by Groups Differing in Feedback Conditions Administered After Initial Testing



			Test 1		Test 2	
Condition	*n*	Measure	*M*	*SD*	*M*	*SD*
F	19	Recall Accuracy	.48	.21	.73	.14
		Response Length[a]	3.14	1.20	3.66	.95
FA	17	Recall Accuracy	.39	.24	.64	.16
		Response Length	2.95	1.21	3.51	.80
FAR	21	Recall Accuracy	.42	.17	.72	.14
		Response Length	3.06	.94	3.57	1.03
NF	19	Recall Accuracy	.52	.19	.52	.19
		Response Length	3.34	1.23	3.02	1.37

Note: F = Feedback only, FA = Feedback and self-assessment activity, FAR = Feedback and self-assessment activity with original response visible, NF = No feedback.

[a]Response length refers to the mean number of words generated in the response to each test item.

Response Productivity

A similar performance pattern among the groups was evident (see Table 1) in the relative lengths of their responses. Participants in the three feedback treatment conditions showed higher response productivity during Test 2—roughly 3.5 words per item—compared to the no-feedback group, which produced an average of 3.02 words per item. Over the course of the two testing sessions, participants in all three feedback groups made comparable increases in response productivity (by an average .5 words per item) while the NF group actually decreased in the number of words produced per test item.

One-way ANOVA of response productivity during Test 2 indicated that these differences were not significant, $F(3, 72) = 1.41$, $p = .25$. However, the Pearson correlation calculated for Test 1 and Test 2 revealed that the scores were significantly related, $r\ (74) = .74, p < .01$. Hence, we controlled for the effect of differences in ability between groups by using Test 1 response productivity as a covariate. Results of the ANCOVA showed the four treatment groups differed significantly, $F(3, 71) = 6.62, p < .01$, $\eta^2 = .22$, in the average number of words generated for each constructed response. A Bonferroni post hoc analysis of the data reported that participants in all three feedback groups wrote significantly longer answers to questions than those who did not receive feedback.

Participant Responses to Feedback

How learners respond to feedback is an important predictor of learning performance in future testing situations (Hattie & Timperley, 2007; Kulhavy & Stock, 1989; Mory, 2004). Hence, we assessed participants' reaction to feedback by examining the time they spent studying feedback, the certainty of their self- assessment, and the mental strategies used to remember story details.

Feedback Study

Computer programs used for all four experimental conditions measured, to the 11th decimal place, the intervening seconds from the moment a short-answer question in test 1 was completed to the instant the onscreen button was clicked to view the next question. Mean times of this latency period per test item (with *SD* in parentheses) for the F, FA, FAR, and NF groups were 4.19 (2.55), 7.25 (4.85), 7.54 (3.90), and 1.46 (0.54) seconds, respectively. Latencies for the NF group were disregarded since this period merely represented the moment needed to submit an answer and request the next question. One-way ANOVA of latency data from the feedback conditions showed significant differences among the three groups; *F*(2, 54) = 25.09,*p* < ,01, η^2 = .48. Results of a Bonferroni post hoc analysis revealed that participants who were asked to simply study the relevant segment from the text took significantly less time on the feedback portion of the test than those who rated their assessment certitude. The time difference between the feedback-only group and those who actively assessed their responses— roughly 80 seconds over the course of test 1—is probably attributable to the additional task demands of those in the FA and FAR groups. The extra time these participants spent during the feedback segments of test 1 also accounted for their overall lengthier experimental sessions compared to those who viewed feedback but did not assess the certainty of their response.

The concept of response certitude (Kulhavy &Stock, 1989) proposes that feedback on incorrect responses is studied longer when such answers are made with high certainty. This lengthier study of feedback should conceivably result in better performance on subsequent testing. For this reason, we were interested in the correspondence between the time students spent examining feedback and their recall performance on the preceding test (test 1) and any difference in improvement from test 1 to test 2.

Separate Pearson correlations calculated for the three feedback conditions between test 1 performance and duration of feedback study reported significant negative correlations for the F (*r* = –,65,*p* < ,01) and FA (*r* = –,51,*p* = .04) groups, but not among FAR participants (*r* = –,29,*p* = .20). Poorer performance on test 1 was associated with significantly longer study of feedback, but not when feedback was accompanied by participants' original response.

Assessment Certitude

Mean certitude ratings for the FA and FAR groups were identical at .52 on a scale of 0 (*totally inaccurate*) to 1 (*totally accurate*) with standard deviations of .20 and .13, respectively. We explored the correlation between participants' certainty on the correctness of their answers in test 1 and their actual performance on this test as well as any improvement in story recall during test 2. Pearson correlations showed a significant positive association between assessment certitude ratings and test 1 recall performance by FA (*r* = .90,*p* < .01) and FAR (*r* = .86, *p* < ,01) participants. Ratings on the certainty of an answer's correctness on test 1 closely paralleled actual performance on the test, regardless of whether one's actual response appeared with the feedback.

A question of greater interest to us, however, was how these ratings corresponded to improved performance on test 2. In this case, the analysis revealed significant negative correlations between assessment certitude ratings and gain scores for both the FA (*r* = –,80,*p* < .01) and FAR (*r* = –,62,*p* < .01) groups. When participants reported low certainty that their answer was correct, they were more likely to improve their performance on the subsequent test compared to those who reported greater certainty about the correctness of their response. We noted a significant negative correlation between assessment certitude and feedback study among those in the FA group (*r* = –,62,*p* < ,01) but not for the FAR group (*r* = –,06,*p* = .79). This observation prompted us to explore the relative association of certitude ratings and feedback study to both test 1 performance and improved recall on test 2.

Relative Impact of Certitude Rating and Feedback Study

We conducted two multiple linear regression analyses—one incorporating test 1 recall and the other using gain scores—separately on FA and FAR data. The two regression models for test 1 data from the FA and FAR groups showed a high goodness-of-fit with nearly 80% of variability in recall accounted for by the combined effect of certitude ratings and feedback study time (adjusted *R*2 values of .79 and .78, respectively). While the regression analyses of gains in recall performance by FA and FAR participants yielded lower coefficients of determination (adjusted *R*2 values of .68 and .34, respectively), results were nevertheless significant for both the former, *F*(2, 14) = 18.12,*p* < .01, and latter, *F*(2, 18) = 6.20,*p* < .01, groups.

The relative influence of assessment certitude ratings and feedback study time can be seen in Table 2. This display shows the standardized regression coefficients (that offer a meaningful comparison between the two different measures) for both variables in test 1 scores and gain scores among the two treatment groups. Assessment certitude proved to be a significant contributor to the regression model for the FA and FAR groups in terms of their performance on test 1 and degree of improved performance on test 2. Regarding test 1 performance, this association was positive whereas assessment certitude was negatively related to gain scores for both groups. This outcome is reasonable since, given that the self assessment was made after test 1, the ratings simply mirrored the raters' perceptions of their performance. By contrast, lower self assessment ratings corresponded to greater overall gains in performance. We speculate that because participants were aware of the need to regulate their learning they presumably did so by studying the feedback provided.

Regression analysis showed the FA and FAR groups differed, however, regarding the influence of feedback study on recall performance. Analysis of test 1 scores by the FAR group, for example, yielded a significant standardized regression coefficient for the feedback study component of the model (β= –,61,*t* = .04), while this was not the case for the FA group (β = .08,*t* = .57).

Comparison of recall performance by the two groups on test 1, their subsequent assessment certitude ratings, and the time spent studying feedback is illustrated by Figure 3 (see top of drawing). This figure depicts a correspondence between test 1 performance and assessment certitude that is nearly identical for the FA and FAR groups. When test 1 responses were presented on the computer screen along with the relevant target text from the story (i.e., the FAR group), the time allocated

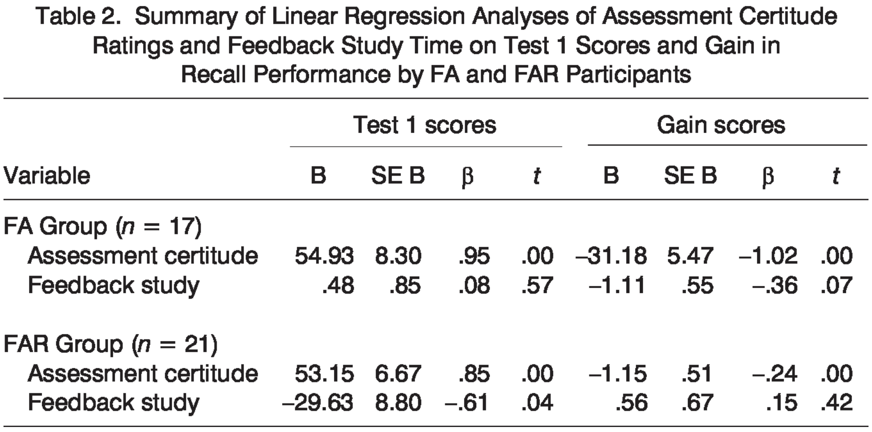


Table 2. Summary of Linear Regression Analyses of Assessment Certitude Ratings and Feedback Study Time on Test 1 Scores and Gain in Recall Performance by FA and FAR Participants

	Test 1 scores				Gain scores			
Variable	B	SE B	β	*t*	B	SE B	β	*t*
FA Group (*n* = 17)								
Assessment certitude	54.93	8.30	.95	.00	–31.18	5.47	–1.02	.00
Feedback study	.48	.85	.08	.57	–1.11	.55	–.36	.07
FAR Group (*n* = 21)								
Assessment certitude	53.15	6.67	.85	.00	–1.15	.51	–.24	.00
Feedback study	–29.63	8.80	–.61	.04	.56	.67	.15	.42

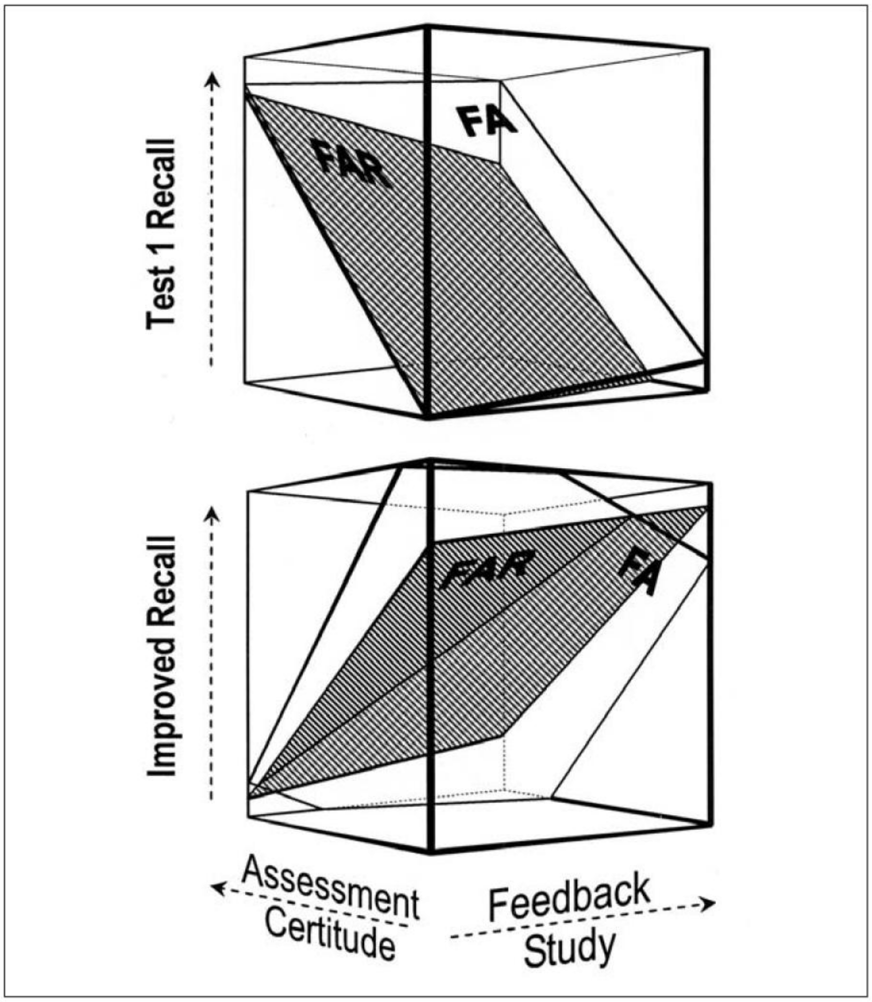


Figure 3. Results of regression analysis comparing the FA and FAR groups on test 1 performance, their assessment certitude ratings, time spent studying feedback, and degree of improved recall on test 2.

for examining feedback was inversely related to recall performance. By contrast, performance on test 1 was relatively unrelated to how long feedback was studied when, as in the FA group, only the target text was available.

Turning to improvement in recall on test 2 (see Table 1), assessment certitude was a significant component in the regression model for both the FA and FAR groups; low assessment certitude ratings were associated with gains in recall performance. On the other hand, standardized regression coefficients for the FA (–,36) and FAR (.15) groups indicated feedback study was a small component of both regression models. Further, when the effect of assessment certitude was controlled, the resulting partial correlation between feedback study and improved performance was moderate for the FA group (*r* = –,47,*t* = .07) and small for the FAR group (*r* =.19,*t* = .42).

Although the correlation between feedback study and improved recall was not significant for either group, considering the magnitude of the correlation obtained given the modest participant pool (*n* = 17) the result for the FA group is nevertheless noteworthy. Further, only FA data exhibited a negative correlation between improved recall and feedback study, suggesting a different response to the feedback provided by the two groups. This difference (Figure 3, bottom) suggests an inverse relationship between feedback study and improved recall by those in the FA group, By contrast, longer feedback study by participants in the FAR group was negligibly but positively related to increases in gain scores.

Analysis of Mental Strategies

At the closure of experimental sessions, participants were prompted to write about mental strategies used to assist them in recalling story details. We speculated that the availability of feedback might alter the approach used to regulate one's own learning. We also thought that ifrating one's assessment certitude heightened the mindfulness of feedback study, participants in the FA and FAR conditions maybe more likely to incorporate feedback in their mental strategies.

Of the 76 participants, 51 persons (67%) reported the use of mental strategies. The remaining 25 participants were classified as having used no mental strategy because they failed to make a response (1 person), stated that no strategy was used (10 persons), or made a response that was irrelevant (14 persons) such as "remembering what I read," Using a latent content analysis strategy (Tashakkori & Teddlie, 1998), four categories of strategies emerged: key feature memorization (38%); feedback iteration (30%); visualization (24%); and personalization (8%). Visualization involved mentally picturing the environment depicted in the story. Personalization strategies consisted of techniques in which participants associated aspects of the story with their personal experiences. Regarding feature memorization strategies, they focused their processing attention on specific key words, places, events, or descriptions from the story. The feedback iteration strategy, which pertained to just the three feedback conditions (FB, FG, and FGR), entailed situations where participants intentionally reviewed and studied the feedback provided to ensure the correctness of future responses.

Figure 4 shows the success in recall performance associated with the various mental strategies reported as well as the frequency of use among all participants in the four experimental groups. For example, among those in the F group, two used personalization, three used visualization, four used feature memorization,

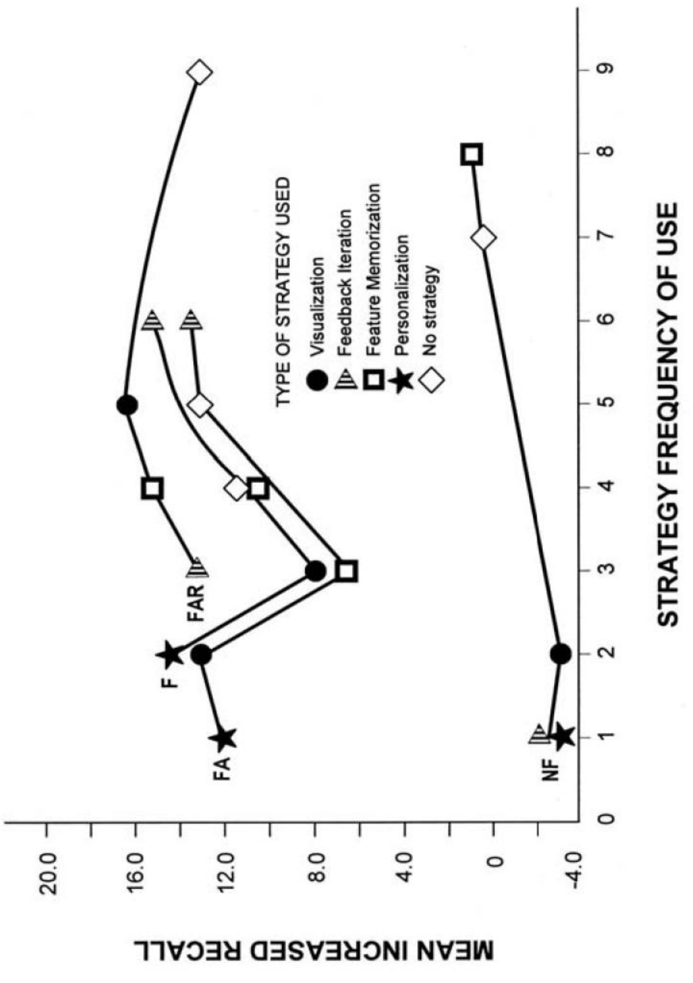


Figure 4. The type and frequency of various mental strategies reportedly used by the four experimental groups as well as the recall performance associated with each strategy.

four listed no strategy, and six used feedback iteration. Memorization of story details was the most frequently used mental strategy reported by those not receiving feedback and was also the method that corresponded to the best recall performance by them. At least twice as many respondents in this group employed this approach than those in the other treatment groups. Among all participants, the personalization strategy was used the least (and not at all by the FAR group).

Among those receiving feedback, but not viewing their original responses (i.e., F and FA groups), the most commonly used mental strategy was an iterative review of the feedback provided. As Figure 4 depicts, this approach is also associated with the greatest improvement in recall by F and FA participants. On the other hand, FAR participants made relatively little use of a feedback iteration strategy but instead were the highest users of visualization among the four treatment groups. The FAR group also contained the highest percentage (43%) of participants categorized as having used no mental strategy.

F and FA groups were identical in the rank order of mental strategies based on user frequency. Both groups were also similar in the relationship between the popularity of a strategy and its corresponding recall performance, an exception being the visualization strategy that yielded high improvements in recall by the few FA group members who employed this approach.

DISCUSSION

This research explored an approach to feedback that might prove useful in cases where formative evaluations cannot be machine-scored, as with multiple- choice exams. In the current study, participants completed constructed-response questions and, after each item, viewed the corresponding target text and rated their performance. Results tentatively support this strategy for providing feedback on constructed responses. Performance significantly improved across all groups that received feedback, even though participants did not receive explicit verification on the correctness of their responses. Given the magnitude of these gains, they are not likely the result of a pretest sensitization effect (Willson &Putnam, 1982), which can also be discounted based on the lack of improvement in recall by the NF group. On the other hand, the improvement in performance following feedback could be explained as simply the result of these groups having not only a second exposure to the story but to the specific paragraph containing the target information. This possibility cannot be discounted since the experiment did not include a treatment group incorporating two exposures to the story but no feedback. Even so, a second reading of the entire story would require more time than simply presenting a second exposure to the target information.

In the FA and FAR groups, the presentation of relevant story material following a question was supplemented by an assessment certitude rating designed to increase the time these participants spent studying the feedback. In both instances, high certitude ratings corresponded with improved performance, particularly among those who saw their original response to a question at the time of rating. A conceivable explanation for this outcome is that participants recognized their poor performance on test 1 and moderated their *cogitative engagement* (Butler & Winne, 1995), scrutinizing the target text shown in the feedback more closely in order to subsequently improve performance (Butler &Winne, 1995; Butterfield & Metcalfe, 2001; Hancock, Stock, & Kulhavy, 1992; Kulhavy &Stock, 1989).

Our data analysis points to several contradictions, though, particularly with respect to the relationship between certitude measures and subsequent feedback study (see Mory, 1994). First, a significant relationship between test 1 performance and feedback study occurred among the F and FA groups with only the latter having assessed the certainty of response accuracy and not for the FAR group. Second, regression analysis of test 1 scores indicated feedback study was a significance component for only the FAR group. Third, analysis of gain scores indicated a significant negative relationship between assessment certitude and feedback study for the FA group only. Finally, regression analysis of gain scores revealed that feedback study contributed relatively little to learning performance.

The FA groups were also more similar to the F group than the FAR group in the types of mental strategies reported. When participants concurrently viewed feedback and their original responses to questions, feedback iteration was the least-used mental strategy, whereas this was the most popular scheme for those who did not have their answers in view. This raises the possibility that those in the FA and F groups retrieved their original responses into working memory, comparing them to the feedback available and mentally rehearsing this information. Since this retrieval activity was not necessary for FAR participants, they were free to consider mental strategies, such as visualization, that use working memory capacity in other ways. This speculation is supported by the fact that a relationship between test 1 performance and feedback study was evident among only the FA and F participants. Theoretically, at least some of the time spent by these participants during the presentation of feedback may have been used to retrieve their original responses. If accurate, this may explain instances where high response certitude is not reflected by correspondingly longer feedback study.

Could potential increases in SRL and learning performance gained through use of assessment certitude in feedback be offset by decreases this may cause in overall instructional efficiency? Such interventions designed to supplement feedback, Mory (2004) has cautioned, may increase the task load for learners and be less efficient than simply providing the correct answer. However, in the current study the increase in overall instructional time among those completing assessment certitude ratings was only negligibly longer less than 2 minutes during a half hour period than those who only received feedback. Such increases in instructional time are acceptable even when compared to the no-feedback group; in this case not providing feedback saved only 3 minutes of instructional time. No differences in the total instructional time were observed between those viewing original responses with feedback and participants who had to mentally recall their responses for comparison with the onscreen feedback. Had the study included a treatment group in which participants received no feedback yet read the entire text twice, an interesting comparison could have been made between a double reading of the text and the use of feedback in which the target text is shown. Conceivably, while both approaches might have achieved the same results, the former may have cost significantly longer instructional time.

Besides this shortcoming, there were other limitations of the study. One was that, while recall performance was calculated using both the accuracy of recall and the degree of elaboration in the answer given, participants were never instructed to construct the most expansive response possible. Another limitation was that participants did not set achievement goals for themselves at the beginning of an experimental session. Goal-setting is a critical component of SRL which, when combined with feedback that reports progress towards one’s goal, has a positive effect on reading comprehension (Schunk, 2003; Zimmerman, 2008b). Rather than merely encouraging participants to “do their best,” future studies on feedback to constructed responses should include the setting of specific learning goals (Schunk, 2003), ideally ones that participants consider personally relevant and meaningful. Finally, the ecological validity of the current research was diminished due to its use of a highly structured but nevertheless inauthentic reading task.

CONCLUSION

There is clearly a need for additional research that can identify feedback schemes for supporting self regulation of performances, such as constructed responses, in which verification of correctness is difficult. Future studies should, in particular, explore the potential interaction between the type of feedback provided and the SRL strategies that participants use. More research is needed, in general, to understand how learners adapt to ongoing instructional feedback by modifying the ways that they self-regulate their learning (Zimmerman, 2008a).

While our research should be considered an initial exploration into types of feedback interventions that may benefit SRL processes, our work nevertheless presents some implications for practice. First, the representation of the same text material from which a constructed response question has been derived can be an effective method for providing feedback, particularly in CAI programs. This approach yielded significant gains in the recall of story text with little cost to instructional overhead, although it remains to be seen whether such improved performance is retained over time. Second, the use of assessment certitude ratings appears to have instructional potential, especially when the initial responses of learners are available for direct comparison with target text. Of particular interest is the possible influence this scheme may have in broadening the repertoire of metacognitive strategies learners use as they monitor their performance and progress during instruction.

As more researchers embark on the study of feedback in circumstances where the possibilities for verification of correctness are limited, additional research questions, variables of interest, and models of feedback will emerge. While the direction such inquiry might take is difficult to predict, the assumption that people use feedback, both external and internal, to regulate and monitor their learning progress is a good place to begin.

REFERENCES

Andre, T., & Thieman, A. (1988). Level of adjunct question, type of feedback, and learning concepts by reading. *Contemporary Educational Psychology, 13*(3), 296-307.
Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., &Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213-238.
Braun, H,I,, Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses

using expert systems. *Journal ofEducational Measurement*, *27*(2), 93-108.

Butler, D. L. &Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review ofEducational Research, 65*(3), 245-281.

Butterfield, B., &Metcalfe, J. (2001). Errors committed with high confidence are hyper- corrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1491-1494.

Campbell, D. T., &Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.

Gay, L. R., & Airasian, P. (2000). *Educational research: Competencies for analysis and application* (6th ed.). Upper Saddle River, NJ: Merrill.

Hancock, T. E., Stock, W. A., & Kulhavy, R. W. (1992). Predicting feedback effects from response-certitude estimates. *Bulletin of the Psychonomic Society, 30*, 173-176. Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Kulhavy, R. W., &Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*(4), 279-308.

Kulhavy, R. W., Yekovich, F. R., &Dyer, J. W. (1976). Feedback and response confidence. *Journal ofEducational Psychology, 68,* 522-528.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning, *Review of Educational Research, 58*(1), 79-97.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389-405,

Mason, B. J., &Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us*. (CLASS Project Research Report No. 9). Lincoln, NE: University of Nebraska-Lincoln, Center for Instructional Innovation.

Moos, D. C., & Azevedo, R. (2008). Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemporary Education Psychology, 33*, 270-298.

Mory, E. H. (1994). The use of response certitude in adaptive feedback: Effects on student performance, feedback study time, and efficiency. *Journal ofEducational Computing Research*, *11*(3), 263-290.

Mory, E. H. (2004). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (2nd ed., pp. 745-783). Mahwah, NJ: Erlbaum.

Reeves, T. C. (2000). Socially responsible educational technology research. *Educational Technology, 40*(6), 19-28.

Schunk, D. H. (1996). Goal and self-evaluative influences during children’s cognitive skill learning, *American Educational Research Journal, 33*(2), 359-382.

Schunk, D. H. (2003). Self-efficacy for reading and writing: Influence of modeling, goal setting, and self-evaluation. *Reading and Writing Quarterly, 19*(2), 159-172.

Schunk, D. H. (2008). *Learning theories: An educational perspective* (5th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Schunk, D. H., & Ertmer, P. A. (1999). Self-regulatory processes during computer skill acquisition: Goal and self-evaluative influences. *Journal of Educational Psychology, 91*(2), 251-260.

Shute, V. J. (2008). Focus on formative feedback. *Review ofEducational Research, 78*(1), 153-189.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology: Combining qualitative and quantitative approaches*. Thousand Oaks, CA: Sage.

Willson, V. L., &Putnam, R. R. (1982). Ameta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal, 19*(2), 249-258.

Zimmerman, B. J. (2002). Achieving self-regulation: The trial and triumph of adolescence. In F. Pajares & T. Urdan (Eds.), *Academic motivation of adolescents* (pp. 1-27). Greenwich, CT: Information Age.

Zimmerman, B. J. (2008a). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166-183.

Zimmerman, B. J. (2008b). Goal setting: A key proactive source of academic self- regulation. In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning* (pp. 267-295). San Diego, CA: Academic Press.