EUROPEAN JOURNAL OF COGNITIVE PSYCHOLOGY
2007, 19 (4/5), 559–579

Psychology Press
Taylor & Francis Group

Improving students' self-evaluation of learning for key concepts in textbook materials

Katherine A. Rawson and John Dunlosky
Kent State University, Kent, OH, USA

Why do students have difficulties judging the correctness of information they recall (e.g., definitions of key concepts in textbooks), and how can students improve their judgement accuracy? To answer these questions, we had college students read six expository passages, each including four key terms with definitions. After reading a text, each key term was presented, and participants (a) attempted to recall the corresponding definition and (b) self-scored the correctness of the response (incorrect, partially correct, or entirely correct). Participants were overconfident, with inflated judgements for responses that were objectively incorrect. When participants could inspect correct definitions while judging their responses, judgement accuracy improved. Counterintuitively, however, some overconfidence remained. We discuss implications of these results for theory, education, and the two questions posed above.

In most educational contexts, students are often presented with large amounts of new material to learn. In primary and secondary grades, current mandates require that students demonstrate competence across many content areas before advancing in school; at the college level, students must learn information from textbooks, lectures, activities, and assignments from several different courses at a time. Given the amount of material that students are expected to learn coupled with a finite amount of time available for study, the task of a successful student is not just to achieve a high level of learning but to regulate study as efficiently as possible.

To regulate study efficiently, a student can profit from accurately evaluating the extent to which they have learned information while studying (Thiede, 1999; Thiede, Anderson, & Therriault, 2003). Consider a student who is studying several different sections in a textbook for an upcoming exam. If the student cannot accurately evaluate how well she has learned a

Correspondence should be addressed to Katherine Rawson, Department of Psychology, Kent State University, PO Box 5190, Kent, OH 44242-0001 USA. E-mail: krawson1@kent.edu

Preparation of this report was supported by Cognition and Student Learning grant No. R305H050038 from the US Department of Education.

http://www.psypress.com/ecp DOI: 10.1080/09541440701326022

given section, the efficiency of study may be compromised. For example, the student may decide she has learned information well when in fact she has not. If, as a result, she prematurely terminates study of that information, she obviously will not achieve an adequate level of learning. Alternatively, the student may decide that she has not learned information well when in fact she has. If she unnecessarily prolongs study of that information, she will limit the amount of time available for studying other information that has not yet been well learned.

Thus, efficient regulation of learning can depend in part on how accurately an individual can evaluate his or her own learning of text materials. Unfortunately, previous research has shown that students' evaluations of their own learning for text materials are only moderately accurate at best. The goal of the present research was to evaluate two hypotheses for why the accuracy of students' evaluations of learning is constrained. In so doing, we will also explore a technique that may improve students' evaluations of learning and ultimately support more efficient regulation of learning. In the next section, we offer a brief historical review of the previous research that led to the two hypotheses to be tested here, and we then describe each of these hypotheses in some detail.

WHAT CONSTRAINS THE ACCURACY OF EVALUATIONS OF LEARNING?

A standard method used in much of the previous research involves an analogue to the study context described for our hypothetical student above. Specifically, individuals are presented with several short texts to study and are then asked to make a judgement for each one. Currently, the most common judgement is what we will refer to as a *global prediction*, in which the individual is asked to predict how well they will do on an upcoming test of the material for each text. We refer to these as *global* because only one prediction is made for an entire text. The accuracy of these predictions is then measured by correlating an individual's predicted test performance with actual test performance across texts. With few exceptions, 20 years of research has shown that the accuracy of global predictions is quite poor, with the mean across individual correlations usually around +.25 (e.g., Maki, 1998; Miesner & Maki, 2007 this issue; Weaver, Bryant, & Burns, 1995; for recent reports of enhanced accuracy, see Dunlosky & Rawson, 2005; Thiede et al., 2003).

In recent research, Dunlosky, Rawson, and Middleton (2005) proposed that global predictions may be relatively inaccurate because of the mismatch between the grain size of the information a student must consider when evaluating learning (i.e., the entire text) and the grain size of the information being tested (i.e., specific definitions, main points, or key concepts from the text). To minimise this mismatch, another kind of judgement has recently been introduced to the standard method, which we refer to here as a *term-specific prediction*. More specifically, Dunlosky, Rawson, and McDonald (2002) had individuals study six short passages, each containing four key terms (e.g., a passage from a nutrition textbook on the body's energy use, which included definitions and explanations of basal metabolism, thermic effects of food, adaptive thermogenesis, and direct calorimetry). After studying a passage, individuals first made a global prediction. They were then presented with each of the four key terms (e.g., basal metabolism) one at a time for a term-specific prediction, in which they were asked to predict how well they would be able to recall the definition of that term on the subsequent test. They were then tested for their memory of the definition for each term. The accuracy of term-specific predictions was measured by correlating an individual's predicted recall with actual recall across all the terms.

The term-specific predictions were expected to be highly accurate, based on the following three assumptions: (a) When presented with the key term for a term-specific prediction, students would self-test using the term as a cue to try to retrieve the target definition, (b) the outcome of this retrieval attempt would be highly diagnostic of subsequent test performance, and (c) students would base their predictions on the outcome of the retrieval attempt. In contrast to our expectation, however, term-specific accuracy was only moderate (mean across individual correlations = .50) and was not significantly greater than the accuracy of the global predictions (mean across individual correlations = .40).

Subsequent work evaluated the three assumptions that led to the (incorrect) expectation that term-specific predictions would be highly accurate. Pertaining to the first assumption, were students self-testing at the time of the term-specific prediction? To answer this question, Dunlosky et al. (2005) used the same method as described above, with one important modification: Prior to making their term-specific prediction, half of the participants were presented with the key term and were explicitly required to overtly recall the target definition (as in Nelson, Narens, & Dunlosky, 2004; Son & Metcalfe, 2005). Students who were forced to attempt recall of target definitions made more accurate term-specific predictions than students who did not overtly attempt recall ($M = 0.73$ vs. $M = 0.57$). Thus, one reason why the accuracy of term-specific predictions can be constrained is that students do not always spontaneously self-test in order to evaluate their learning (cf. Kelemen, 2000).

Note, however, that the accuracy of term-specific predictions was still less than perfect even when students were forced to attempt recall of the target definitions. With respect to the second assumption, was accuracy constrained because the outcome of the retrieval attempt was not highly diagnostic of subsequent recall? Follow-up analyses showed that the mean intraindividual correlation between prejudgement recall and criterion recall across terms was around .90, suggesting that the outcome of prejudgement recall was a highly diagnostic cue for predicting criterion recall. Why then was the accuracy of term-specific predictions still constrained? Consider again the third assumption, which states that students base their predictions on the outcome of the retrieval attempt. The mean intraindividual correlation between prejudgement recall and term-specific predictions was around .67. Thus, although the outcome of prejudgement recall was diagnostic of subsequent test performance (.90), students apparently did not fully capitalise on this cue when predicting test performance.

Most important for present purposes, a follow-up experiment indicated that students were not adequately evaluating the correctness of the outcome of their retrieval attempts. In this experiment, all students performed prejudgement recall. Importantly, some participants were then asked to make a *self-score judgement*. For this judgement, they were explicitly asked to score the correctness of their recall response, using the following prompt: "If the correctness of the definition you just wrote was being graded, do you think you would receive no credit, partial credit, or full credit?" In evaluating their performance, students often assigned full credit to responses that were only partially correct, partial credit to responses that were completely correct, and most troublesome, they frequently assigned partial or even full credit to responses that were completely incorrect. The present research was designed to explore the inaccuracy of these self-score judgements and the difficulties students have in evaluating the correctness of their own responses when self-testing memory for target information.

WHY ARE SELF-SCORE JUDGEMENTS INACCURATE?

As argued above, students must be able to identify material that they have not learned well enough so that subsequent study can be focused on that content. One strategy for assessing how well something has been learned is to self-test—indeed, the original expectation was that term-specific predictions would be quite accurate because they afforded an opportunity to self-test memory for information at an appropriate grain size for evaluation. However, the value of self-testing for guiding subsequent regulation of study hinges critically on the extent to which an individual can accurately evaluate the *correctness* of the outcome of the self-test, which leads us to our focal question: What factors constrain the accuracy of students' self-evaluations of target information recalled during self-testing of memory for expository text content?

According to the *absence of standard hypothesis*, when students do not have access to the objectively correct target information—i.e., the objective standard of evaluation—they will have difficulties in evaluating the correctness of recalled information. Such difficulty seems inevitable, because if a student retrieves incorrect information, he or she presumably has not adequately learned the actual sought-after target information, so comparing the retrieved information to this sought-after information would be impossible without external support (for an elaboration of this idea, see Hacker, 1998). Without access to an external standard, students may turn to other cues to evaluate the adequacy of the retrieved information, such as the quantity of information recalled during the retrieval attempt, which may not be highly related to the objective correctness of the retrieved information (e.g., Baker & Dunlosky, 2006; Morris, 1990). According to this hypothesis, however, students may be capable of better evaluating the correctness of their recall output if they have access to an external standard. Examining the degree of consistency between a retrieved response and the objectively correct response may serve as a means to more accurately evaluate the correctness of the retrieved response.

The *limited competence hypothesis* states that students have limited ability to evaluate the correctness of retrieved information, even if an external standard is available for comparison. According to this hypothesis, providing students with an external standard for comparison to their generated responses will not significantly improve the accuracy of self-score judgements. The plausibility of this account is suggested by several studies in the text comprehension literature. Research on error detection has shown that students are often unable to identify factual inconsistency between pieces of information within a text. For example, Otero and Kintsch (1992) presented readers with short texts, some of which contained two sentences that were inconsistent with one another (e.g., "Superconductivity ... has only been obtained by cooling certain materials to low temperatures near absolute zero", and then later, "[u]ntil now superconductivity has been achieved by considerably increasing the temperature of certain materials"). Forty per cent of the inconsistencies went undetected. Similarly, Johnson and Seifert (1994) presented readers with fictional police reports about a warehouse fire, in which one of the earlier reports was subsequently updated (e.g., "... they have reports that cans of oil paint and pressurised gas cylinders had been present in the closet before the fire", and later, "... the closet reportedly containing cans of oil paint and gas cylinders had actually been empty before the fire"). When subsequently tested for understanding of the cause of the fire, over 90% of the participants made at least one direct, unqualified reference to the volatile materials. Studies such as these suggest that students may have difficulty recognising the inconsistency between two pieces of explicitly stated text information. Even more relevant, other research suggests that students have difficulty recognising inconsistencies between generated responses and explicit text information. Howe (1970) had readers listen to a short passage, recall it, and then listen to it again. In each of the following 3 weeks, they first recalled the text and then listened to it again. Information that was incorrectly recalled on initial tests continued to be recalled on subsequent tests despite re-presentation of the text (in fact, inclusion of previously recalled incorrect information was two to three times more likely than inclusion of previously unrecalled correct information). This finding is suggestive of individuals' inability to recognise the inconsistency between what they recall and explicitly presented information that is objectively correct.

GOALS OF THE PRESENT RESEARCH

The present research was designed to evaluate the absence of standard hypothesis and the limited competence hypothesis. For this purpose, we adapted the method used by Dunlosky et al. (2005) described above. Participants were presented with several short texts each containing four key terms. After reading a text, readers were presented with each of the four key terms one at a time and were asked to recall the definition of the term. After the recall attempt, all participants were asked to make a self-score judgement in which they rated the correctness of the generated response. For one group of participants, the correct definition was presented along with the participant's generated response at the time of the self-score judgement (hereafter referred to as the *standard* group). For the other group, only the generated response was presented (hereafter referred to as the *no standard* group).

The absence of standard hypothesis predicts that individuals who are presented with the correct answer will make more accurate self-score judgements than individuals who are not shown the correct answer. Specifically, individuals in the standard group will be more likely to assign generated responses to the appropriate categories ("no credit" for incorrect responses, "partial credit" for partially correct responses, and "full credit" for correct responses). In contrast, the limited competence hypothesis predicts that the accuracy of the self-score judgements will not significantly differ for the standard and no standard groups. Note, of course, that these two hypotheses are not mutually exclusive, and both mechanisms may undermine the accuracy of people's self-assessments. Thus, providing an external standard may improve accuracy, yet students' self-score judgements may still demonstrate some biases even with a standard. Importantly, this possibility can be explored using the methods adopted in the present research.

METHOD

Participants and design

Fifty-six undergraduates participated to partially satisfy a course requirement in Introductory Psychology. Participants were randomly assigned to one of two groups, standard ($n = 30$) or no standard ($n = 26$).

Materials

The materials were the same as those used by Dunlosky et al. (2005) and included seven expository texts (one sample and six critical) that were taken from introductory-level textbooks from various undergraduate courses (e.g., nutrition, family studies, communication). Texts were between 271 and 281 words long, with Flesch-Kincaid scores ranging from grade levels 10 to 12. Each text contained four key terms (presented in capital letters), and each term was immediately followed by a one-sentence definition (e.g., "ADAPTIVE THERMOGENESIS refers to when the body expends energy to produce heat in response to a cold environment or as a result of overfeeding"). Macintosh computers presented all materials and recorded all responses.

Procedure

Participants were given detailed instructions about each phase of the task. Before beginning the critical study trials, participants practised each task with the sample text and test questions to familiarise them with the kind of text and tests they would receive in the critical trials.

The critical texts were presented in random order for each participant. Each text was presented individually in its entirety for self-paced study. Participants clicked a button on the screen to indicate when they were done studying a text. Immediately after reading a given text, participants were asked to make a global prediction with the following prompt: "How well will you be able to complete a test over this material? 0 = definitely won't be able, 20 = 20% sure I will be able, 40 = 40% sure ... 100 = definitely will be able." After making the global prediction, participants were presented with the following prompt: "Please practice recalling the following information from the text you just read", followed by one of the key terms from the text (e.g., "Define: adaptive thermogenesis"). Participants typed their response into a text field on the screen. After participants indicated they were done recalling the definition by clicking on a button, they were asked to evaluate or "self-score" the response they generated. For participants in both groups, the participant's response appeared on the screen along with the following prompt: "If the correctness of the definition you just wrote was being graded, do you think you would receive no credit, partial credit, or full credit?" For participants in the standard group, the correct definition of the term was also presented above the participant's response. After self-scoring their response, the response (and definition) were removed from the screen. Participants were then asked to predict how well they would be able to define that term on the actual test, using the 0–100 scale described above.

Participants completed this procedure for each of the four terms in a text (prejudgement recall, self-score judgement, and term-specific prediction), and then the criterion test was administered for that text. For the criterion test, each of the four terms was presented individually on the screen, and the participants typed their recall of the definition into a text field. After the criterion test for a text had been completed, participants studied the next text and so on until the procedure had been completed for all six texts.

RESULTS

Self-score judgements

For each participant, we computed a mean across self-score judgements for all items (assigning a value of 0 to a self-score judgement of "no credit", 50 to a judgement of "partial credit", and 100 to a judgement of "full credit"). We then computed a mean across the individual means in each group. Overall, self-score judgements were significantly lower for the standard group ($M = 34$, $SE = 3$) than for the no standard group ($M = 44$, $SE = 3$), $t(54) = 2.25$, $p < .05$.[1]

Of greater interest is examination of the self-score judgements for each kind of prejudgement recall response. Each prejudgement recall response was scored using a gist criterion. Specifically, each definition consisted of two to four idea units, and an idea unit was counted as being present in a participant's response when the participant either stated the idea verbatim or correctly paraphrased the original text. On the basis of this scoring, we then assigned each prejudgement recall response to one of five categories, which were chosen because they reflect the kinds of qualitative outcomes that should be reflected in people's self-score judgements (see Dunlosky et al., 2005). These categories were: omission error (no response), commission error (a completely incorrect response), partially correct (a response

[1] For all analyses involving self-score judgements, we also conducted nonparametric inferential tests (Mann-Whitey Test) to compare self-score judgements from the standard group versus the no standard group. All of these tests supported the same statistical conclusions as obtained from the parametric tests.

containing at least one correct idea unit), partial plus commission (a response that contained incorrect information along with at least one correct idea unit), and correct (a completely accurate response). Partial plus commission responses were rare (on average, less than one per participant), so we do not consider this response category further. For each group, the percentage of total responses in each category is indicated above each bar in Figure 1.

For each individual, we computed the mean self-score judgement for responses within each of the four categories of interest. The means across these individual values in each response category for each group are depicted by the bars in Figure 1. A 2 (group) × 4 (response category) mixed factor analysis of variance (ANOVA) revealed significant main effects of group and response category, $F(1, 41) = 9.04$, $MSE = 450.67$, $p < .01$, and $F(3, 123) = 176.66$, $MSE = 253.76$, $p < .001$, respectively, and a significant interaction, $F(3, 123) = 12.10$, $MSE = 253.76$, $p < .001$. Follow-up tests revealed that the magnitude of self-score judgements for the two groups did not significantly differ for omissions, partially correct responses, or correct responses, all ts ≤ 1.20, ps $> .10$. In contrast, self-score judgements for commissions were significantly lower in the standard group than in the no standard group, $t(54) = 5.92$, $p < .001$. Thus, when individuals were provided standards for evaluating their responses, they were better able to judge the correctness of those responses, consistent with the absence of standard hypothesis. This improvement in the absolute accuracy of the self-score judgements for commissions may be particularly important, considering that commissions were the most frequent kind of response—across participants, 38% of all prejudgement recall responses were commissions (compared to 29% omissions, 16% partially correct responses, and 14% correct responses).

Note, however, that although providing standards improved the accuracy of self-score judgements for commissions, students still inappropriately awarded credit to some responses that were completely incorrect, which provides some support for the limited competence hypothesis. To examine this pattern of errors further, for each individual, we computed the percentage of commissions that were assigned to each of the three self-score judgements (no credit, partial credit, or full credit). We then computed the mean percentage across participants for each judgement category in each group. These means are reported in Table 1. Surprisingly, even when a standard for evaluation was explicitly provided, students still judged incorrect responses as partially or fully correct 43% of the time.

Why are students sometimes unable to recognise the inconsistency between generated responses and explicit standards for evaluation? One possibility is that self-score judgements may be partly influenced by the length of the response. For instance, students may believe that if they

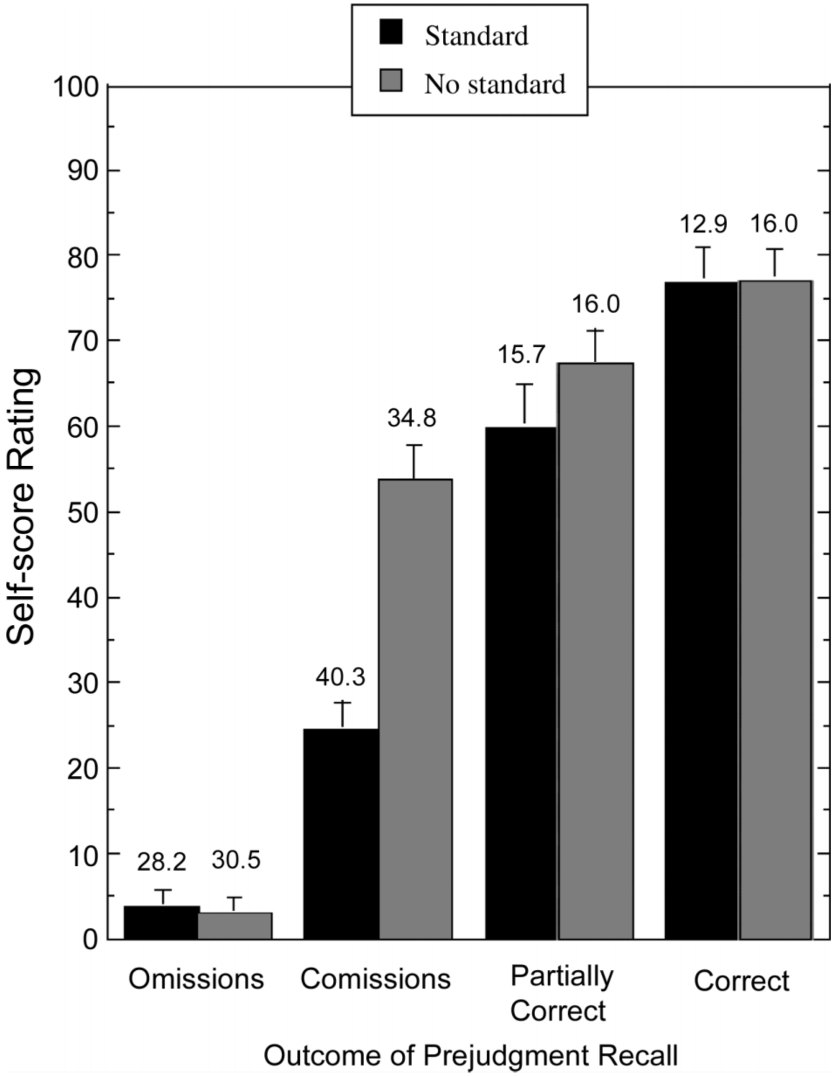

Figure 1. Bars represent mean self-score judgement magnitude for each of four kinds of response as a function of group. Error bars are standard errors of the mean. Percentages above each bar indicate the percentage of responses (out of 24) assigned to each response category depicted in graph (values do not total to 100 because one low-frequency response category was omitted from analysis; see text for details).

produce enough, something in what they produce must be right. This possibility is consistent with the accessibility hypothesis of metacognitive judgements, which states that judgements reflect the amount and speed with

TABLE 1
Percentage of commissions assigned to each of three self-score judgment categories

	No credit		*Partial credit*		*Full credit*	
	M	*SE*	*M*	*SE*	*M*	*SE*
Standard	*57*	*5*	*37*	*5*	*6*	*2*
No standard	17	4	58	4	25	5

SE = standard error of the mean.

which to-be-evaluated information is accessed from memory (e.g., Benjamin, Bjork, & Schwartz, 1998; Koriat, 1993; Morris, 1990). To explore this possibility, for the terms that elicited commission errors, we correlated the number of words in each generated response with the self-score judgements. Whereas response length was significantly correlated with self-score judgements in the no standard group ($r = .29$, $p < .01$), response length was not significantly correlated with self-score judgements in the standard group ($r = .10$, $p = .09$). Thus, the possible effects of response quantity did not strongly bias students' judgements, especially for those who had access to the objective standards while evaluating the correctness of their recall.

Another possibility is that although an explicit standard was presented on every trial, students may not always have referred to the standard when making their self-score judgement. One indicator of whether they referred to the standard can be inferred from how much time participants took to make the self-score judgements (referred to hereafter as *self-score time*). For each individual in each group, we computed (a) the mean self-score time for commission errors that were assigned a self-score judgement of no credit and (b) the mean self-score time for commission errors that were assigned a self-score judgement of either partial credit or full credit. (We combined the latter two judgement categories to reduce the number of participants who would have been dropped from analysis because they did not have any commission errors in one of the categories.) Consider first the self-score times for participants in the no standard group, as these results provide a baseline for when participants could not consult an external standard. Across participants in the no standard group, mean self-score time was 2.5 s ($SE = 0.6$) for commission errors that were assigned no credit and 2.3 s ($SE = 0.2$) for commission errors that were assigned partial or full credit. By comparison, across participants in the standard group, mean self-score time was 7.9 s ($SE = 0.7$) for commission errors that were assigned no credit and 6.9 s ($SE = 0.8$) for commission errors that were assigned partial or full credit. Self-score times were consistently longer in the standard group than in the no standard group, ts > 4.90, suggesting that participants in the standard group were consulting the explicit standard when making their self-score judgements. More important, self-score judgement times in the standard group did not significantly differ for commissions assigned no credit and commissions assigned partial or full credit, $t(21) = 1.04$. Although tentative, these results provide further support for the limited competence hypothesis because students were apparently consulting the external standard but still failed to recognise inconsistencies between recalled information and objectively correct information some of the time.

Cued recall performance on the prejudgement test and the criterion test

For prejudgement recall and for criterion test recall, each cued recall response was scored using a gist criterion. Each definition consisted of two to four idea units, and an idea unit was counted as being present in a response if the idea unit was stated verbatim or if it was a correct paraphrase of the original text. A response was then assigned a score of 0% if it contained no correct idea units, 50% if it contained at least 50% but less than 100% of the idea units contained in the definition, and 100% if it contained all of the idea units in the definition. For each participant, we then computed mean recall performance across all individual response scores on prejudgement recall and mean recall performance across all individual response scores on the criterion test; thus, individual mean values could range from 0 to 100%. Mean recall across participants in each group for prejudgement recall and the criterion test are reported in Table 2.

A 2 (group) × 2 (time of test: prejudgement recall vs. criterion test) mixed factor ANOVA yielded a significant main effect of time of test and a significant interaction, $F(1, 54) = 59.46$, $MSE = 58.76$, $p < .001$, and $F(1, 54) = 41.50$, $p < .001$, respectively. The main effect of group was not significant, $F < 1.85$. Follow-up tests revealed that whereas performance significantly improved from prejudgement recall to criterion test for the standard group, performance did not significantly improve in the no standard

TABLE 2
Performance on pre-judgment recall and criterion recall

	Standard		*No standard*	
	M	*SE*	*M*	*SE*
Prejudgment recall	20	3	24	3
Criterion recall	41	4	25	4

Performance is reported as a percentage. *SE* = standard error of the mean.

TABLE 3
Performance on criterion recall test as a function of outcome on the prejudgment recall test

	Standard		*No standard*	
Outcome of prejudgment recall	*M*	*SE*	*M*	*SE*
Omission	16.5	3.5	2.3	1.0
Commission	35.6	4.8	8.8	3.0
Partially correct	58.2	5.0	35.8	4.2
Correct	91.8	2.7	84.7	5.6

Performance is reported as a percentage. SE = standard error of the mean.

group, $t(29) = 8.22$, $p < .001$, and $t(25) = 1.56$, respectively. Moreover, whereas prejudgement recall in the two groups did not significantly differ, criterion test performance was significantly greater in the standard group than in the no standard group, $t(54) = 0.88$ and $t(54) = 3.00$, $p < .01$, respectively.

To further explore the effects of providing an external standard on criterion test performance, we examined criterion test performance as a function of prejudgement recall status. Specifically, for each individual, we computed mean criterion recall conditionalised on the outcomes of prejudgement recall, namely, for omission errors, commission errors, partially correct responses, and correct responses. Means across individual values for each group are reported in Table 3. Criterion recall was greater in the standard group than in the no standard group for prejudgement omissions, commissions, and partially correct responses, ts > 3.36. Criterion recall for prejudgement correct responses in the two groups did not significantly differ, $t(46) = 1.16$.

Within the standard group, recall significantly improved from prejudgement to criterion test for omissions, commissions, and partially correct responses, ts > 3.45, whereas recall significantly declined from prejudgement to criterion test for correct responses, $t(23) = 2.88$, $p < .01$.[2] Within the no

[2] By definition, mean prejudgement recall was 0% for omissions and commissions, and 100% for correct responses. Mean prejudgement recall across partially correct responses was 40% (SE = 2) in the standard group and 36% (SE = 3) in the no standard group. Note that prejudgement recall for partially correct responses is less than 50% because of the differences between the criterion used to score recall and the criterion used to categorise prejudgement responses. A response was categorised as partially correct if it contained at least one of the idea units from the correct answer, whereas a recall score of half credit was assigned only if the response had 50% or more of the correct idea units. Thus, some responses were categorised as partially correct but received scores of 0 (e.g., when only one of four idea units was recalled). The logic of this categorisation scheme was to keep the commission category "pure", so that responses in this category contained no objectively correct information. Accordingly, responses that contained any correct information were included in the partial category, even if it did not contain enough to merit a half credit score.

standard group, recall significantly improved from prejudgement to criterion test for omissions and commissions, $t\text{s} > 2.24$, whereas recall significantly declined from prejudgement to criterion test for correct responses, $t(23) = 2.74$, $p < .05$. Recall for prejudgement and criterion test did not significantly differ for partially correct responses, $t(25) = 0.17$. Importantly, although performance gains in some categories were observed in both groups, greater gains were obtained in the standard group. In addition to the significant main effect of time of test, the interaction of time of test and group was significant for omissions, $F(1, 50) = 14.12$, $p < .001$, commissions, $F(1, 54) = 20.51$, $p < .001$, and partially correct responses, $F(1, 53) = 8.84$, $p < .01$; the interaction for correct responses was not significant, $F = 1.53$. Performance on the criterion test was greater for the standard group than for the no standard group for omissions, commissions, and partially correct responses, $t\text{s} > 3.41$.

These results show that providing an external standard not only improves the accuracy of self-score judgements, but it can also improve memory for the evaluated information. Moreover, just as the effects of providing a standard on self-score judgement accuracy depended on the kind of prejudgement recall response, the effects of providing a standard on memory also depended on the kind of prejudgement recall response. The effects of providing a standard on memory demonstrated here also contribute to the literature investigating the effects of retrieval practice and restudy on recall. We will consider these issues further in the Discussion.

Global and term-specific prediction accuracy

Although prediction accuracy is not relevant to empirically evaluating the focal hypotheses of this research, for interested readers and for archival purposes, we briefly report conventional analyses of prediction accuracy here. For each individual, we computed a gamma correlation between the six global predictions and criterion recall for each text. Means across individuals' correlations did not differ significantly between the group receiving the standard ($M = 0.36$, $SE = 0.09$) and the no standard group ($M = 0.21$, $SE = 0.13$), $t(52) = 0.99$.

For the term-specific predictions, we computed a gamma correlation across the 24 term-specific predictions and criterion recall for the corresponding terms for each individual. Means across individual correlations were reliably lower for the standard group ($M = 0.52$, $SE = 0.06$) than for the no standard group ($M = 0.68$, $SE = 0.05$), $t(53) = 2.01$. Although one may expect that predictive accuracy would improve with the presence of feedback, the feedback itself influenced criterion recall (as discussed above), which in turn would undermine predictive accuracy (see also, Kimball & Metcalfe, 2003). Perhaps most important, note that accuracy was moderate-to-high for term-specific predictions and higher than the accuracy of the global predictions, which replicates outcomes reported by Dunlosky et al. (2005).

DISCUSSION

The present study was motivated by previous research showing that when students are asked to assess their learning of key terms from expository text, the accuracy of these judgements is constrained. One way students could assess their learning is by attempting to recall the target definition and then by evaluating the *correctness* of the output of the retrieval attempt. The idea here is that this judgement involves a participant's evaluation of how well the retrieved information represents the actual meaning of the definition. In this way, these judgements tap metacomprehension because they involve evaluating one's understanding of a concept. Unfortunately, students' judgements about the correctness of retrieved definitions are only moderately accurate (Dunlosky et al., 2005), which suggests that judgement accuracy is partly constrained by inaccurate metacomprehension. In the present research, we tested two specific hypotheses for the limited accuracy of people's evaluations of the correctness of retrieved information—the limited competence hypothesis and the absence of standard hypothesis.

According to the limited competence hypothesis, students have limited ability to recognise inconsistencies between information they have recalled and objectively correct information, even when provided with an external standard for comparison. Results provided some support for this hypothesis. In particular, providing an external standard did not significantly improve the accuracy of self-score judgements for prejudgement recall responses that were partially or completely correct. Furthermore, although the provision of standards did improve judgements for commission errors somewhat, students who were provided an external standard still assigned partial or even full credit to 43% of responses that were completely incorrect.

Why might students be limited in the extent to which they can recognise inconsistencies between information they have recalled and objectively correct information? Two possibilities explored here were (a) that students' evaluations are unduly influenced by response quantity, and (b) that they did not consult the objectively correct information when evaluating their response. Secondary analyses suggested that the influence of both factors was minimal. Although not evaluated here, another plausible explanation for at least some of the inflated self-scores may be that students expect to receive some credit just for trying, based on prior classroom experiences involving similar evaluation situations. Although this explanation is unlikely to account for cases in which students awarded full credit to completely incorrect responses, it may account for some cases in which incorrect responses are assigned partial credit. Future research could evaluate this account by providing participants with explicit training and examples of how response correctness will be graded on the final test.

Another possibility is that the extent to which students can recognise inconsistencies between their responses and correct information is constrained by limited working-memory capacity. According to the construction-integration theory of text comprehension (Kintsch, 1988, 1998), text material is processed within a limited capacity processing system. Thus, text material is processed in cycles, with the amount of input processed in a cycle roughly equivalent to a simple sentence. Only a limited amount of information may be carried over from one processing cycle to the next, because some capacity must be available for processing the subsequent input to the processing system. Similarly, in order to recognise an inconsistency between a response and a standard, an individual must have both pieces of information concurrently active in working memory (Otero & Kintsch, 1992). Applied to the present task, students may not be able to carry over all of the content in their response to the next cycle in which the objectively correct information is processed (or vice versa). Although external presentations of one's response and the correct standard may reduce the capacity demands of making a comparison, visual comparison of the two pieces of information is not enough. That is, even though they may contain some of the same surface information (e.g., words or syntax), such overlap does not ensure that they will contain the same semantic information. Thus, the semantic content of the response and the correct standard must be processed together, which may exceed the capacity of the working memory system in which the processing of semantic information takes place.

Importantly, in addition to the evidence for the limited competence hypothesis, we also found some support for the absence of standard hypothesis. According to this hypothesis, evaluation of the correctness of recall output will be constrained when students do not have an external standard to which their output can be compared. Consistent with this hypothesis, providing an external standard improved the accuracy of self-score judgements. Although students in the standard group were still inappropriately assigning some credit to incorrect responses 43% of the time, this bias was significantly less than the 83% error rate observed in the no standard group (Table 1).

Effects of standards on memory

In addition to improving the accuracy of self-score judgements, providing an external standard also improved memory for the target information.

The finding that criterion recall was greater in the standard group than in the no standard group is consistent with previous research on the effects of retrieval practice and restudy. For example, Cull (2000) presented individuals with vocabulary word pairs (e.g., "handsel–payment") for an initial study trial. The word pairs were then presented three more times in one of three conditions: Individuals either attempted to retrieve the second word of a pair when presented with the first word as a cue, restudied the word pair, or both (i.e., a retrieval attempt followed by presentation of the correct answer for restudy). Performance on a final cued recall test was greater after retrieval plus restudy than after retrieval alone or restudy alone.

Consistent with the current results, recent research has further suggested that the efficacy of restudy after retrieval for improving memory depends on the nature of the initial retrieval response. Pashler, Cepeda, Wixted, and Rohrer (2005) presented individuals with Luganda–English translation equivalents (e.g., "leero–today") for two study trials. Cued recall for the vocabulary pairs was then tested twice. One group did not receive feedback about the accuracy of their responses, whereas a second group was provided with the correct answer after each retrieval attempt. Recall on the second test was greater when correct answers had been provided during the first test than when no answers were provided. However, this effect was only significant for responses that were omissions or commissions on the first test; the two groups did not significantly differ in criterion recall for initially correct responses. These results mirror the present findings: Criterion recall was greater in the standard group than in the no standard group, but only for those items that were not correctly recalled during prejudgement recall. In this way, the present research provides an important advance beyond previous research on retrieval practice and restudy, by showing that these effects generalise from simple verbal materials (word pairs) to more complex text materials. Our results converge with those reported by Kang, McDermott, and Roediger (2007 this issue), who also found that feedback provided after short answer practice tests was particularly beneficial for initially incorrect responses. However, we should note that Kang et al. did not find this effect of feedback following multiple choice practice tests. Furthermore, Butler and Roediger (2007 this issue) did not find any effect of feedback for either kind of practice test, although this may have been due to a relatively high rate of initially correct responses. In short, further exploring the effects of feedback on memory for more complex materials is an important direction for future research.

Educational implications

What are the potential implications of these results for promoting the effectiveness of students' self-regulated learning? Given that self-testing is often implicitly or explicitly recommended to students, the present research suggests that self-testing may be an effective strategy for improving students' self-regulated learning when a standard for evaluating the outcome of the self-test is available. First, providing a standard that consists of the correct answer can improve memory for the to-be-learned content. Second, providing a standard also improves the accuracy of students' evaluations of their own learning, which may further improve learning by supporting effective self-regulation of subsequent study. To effectively regulate study, students must be able to accurately identify material that they have not learned well enough so that subsequent study can be focused on that content. Of course, we are not suggesting that accurate monitoring is sufficient for effective self-regulation, because students must also be motivated to study and appropriately use the output from monitoring to guide study. Nevertheless, the finding that providing a standard improves the accuracy of self-score judgements suggests that it could enhance the effectiveness of self-regulated study. Consider again the results presented in Table 1. Assuming that students would choose to restudy any item for which their self-test response was partially or completely incorrect, students who were provided with a standard would only have failed to select 6% of their incorrect responses for restudy. Without a standard, 25% of the incorrect responses would not have been studied further.

The present results also have implications for instruction. Textbooks often contain end-of-chapter review questions (e.g., a list of the key concepts covered in that chapter), and instructors often provide students with study guides to prepare for exams. The implicit or explicit instruction to students is that they should use these materials as a basis for self-testing. Our previous research suggests that students may not always self-test their memory even when provided with cues that afford retrieval practice. Instead of evaluating memory using a full-blown retrieval attempt, they may assume they know the answer if they feel familiar with the terms themselves (e.g., Reder & Ritter, 1992). The present research further suggests that even when students do self-test, the effectiveness of the strategy for evaluating and improving learning will depend critically on whether an external standard is provided for comparison to the outcome of the self-test. Unfortunately, although many textbooks provide key term lists at the end of chapters, they often do not provide the correct responses in a concise and easy-to-reference manner. Of course, students could look back through the chapter to find each definition to check against their self-test response, but this would obviously be cumbersome and time consuming. Likewise, study guides provided by instructors are rarely accompanied by corresponding answers. Instructors may assume that students will revisit lecture notes to find the correct information for comparison to self-test responses. However, students' lecture notes are often grossly incomplete and lacking in organisation (e.g., Titsworth & Kiewra, 2004), and thus it will often be difficult or impossible for students to locate standards on their own. Accordingly, one way in which instructors can support more effective student self-regulated learning is to provide external standards for use during self-testing.

Although we suggest that self-testing followed by an external standard for evaluating the outcome can be an effective strategy for improving students' self-regulated learning, we do not mean to imply that the self-testing strategy is sufficient for meeting learning goals in all situations. Indeed, for much of the content in many courses, the expectation is that students will be able to think critically about and apply the information they learn. Accurate evaluation of one's self-testing outcomes against an external standard can improve memory for that information and inform subsequent restudy decisions to further improve learning, but it does not guarantee that students will be able to think critically about or apply what they are recalling. Nevertheless, the importance of being able to remember key terms and concepts should not be diminished: Students will not be able to apply a concept they cannot remember, nor think critically about those they cannot recall. Thus, the self-testing strategy may benefit students' comprehension of key concepts at least indirectly and allow them to more efficiently meet their learning goals.

REFERENCES

Baker, J., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin and Review*, *13*, 60–65.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*, 55–68.

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*, 514–527.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235.

Dunlosky, J., & Rawson, K. A. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes*, *40*, 37–55.

Dunlosky, J., Rawson, K. A., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. Perfect & B. Schwartz (Eds.), *Applied metacognition* (pp. 68–92). Cambridge, UK: Cambridge University Press.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language*, *52*, 551–565.

Hacker, D. (1998). Self-regulated comprehension during normal reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 165–191). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Howe, M. J. A. (1970). Repeated presentation and recall of meaningful prose. *Journal of Educational Psychology*, *61*, 214–219.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1420–1436.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.

Kelemen, W. L. (2000). Metamemory cues and monitoring accuracy: Judging what you know and what you will know. *Journal of Educational Psychology*, *92*, 800–810.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory and Cognition*, *31*, 918–929.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*, 163–182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*, 609–639.

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Miesner, M. T., & Maki, R. H. (2007). The role of test anxiety in absolute and relative metacomprehension accuracy. *European Journal of Cognitive Psychology*, *19*, 650–670.

Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 223–232.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised method for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Method*, *9*, 53–69.

Otero, J., & Kintsch, W. (1992). Failures to detect contradictions in a text: What readers believe versus what they read. *Psychological Science*, *3*, 229–235.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3–8.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 435–451.

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory and Cognition*, *33*, 1116–1129.

Thiede, K. W. (1999). The importance of monitoring and self-regulation during multi-trial learning. *Psychonomic Bulletin and Review*, *6*, 662–667.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of text. *Journal of Educational Psychology*, *95*, 66–73.

Titsworth, B. S., & Kiewra, K. A. (2004). Spoken organizational lecture cues and student notetaking as facilitators of student learning. *Contemporary Educational Psychology*, *29*, 447–461.

Weaver, C. A., Bryant, D. S., & Burns, K. D. (1995). Comprehension monitoring: Extensions of the Kinstch and van Dijk model. In C. A. Weaver, S. Mannes, & C. R. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 177–193). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.