Received: 10 June 2023 | Accepted: 30 November 2023

DOI: 10.1111/bjep.12651

INVITED ARTICLE

Learning from errors versus explicit instruction in preparation for a test that counts

Janet Metcalfe[1] | Judy Xu[1] | Matti Vuorre[2] | Robert Siegler[1] | Dylan Wiliam[3] | Robert A. Bjork[4]

[1]Columbia University, New York, New York, USA

[2]Tilburg University, Tilburg, The Netherlands

[3]University College London, London, UK

[4]UCLA, Los Angeles, California, USA

Correspondence
Janet Metcalfe, Department of Psychology, Columbia University, New York, NY 10027, USA.
Email: jm348@columbia.edu

Funding information
U.S. Department of Education; United States Department of Education Institute of Education Science, Grant/Award Number: R305A150467

Abstract

Background: Although the generation of errors has been thought, traditionally, to impair learning, recent studies indicate that, under particular feedback conditions, the commission of errors may have a beneficial effect.

Aims: This study investigates the teaching strategies that facilitate learning from errors.

Materials and Methods: This 2-year study, involving two cohorts of ~88 students each, contrasted a learning-from-errors (LFE) with an explicit instruction (EI) teaching strategy in a multi-session implementation directed at improving student performance on the high-stakes New York State Algebra 1 Regents examination. In the LFE condition, instead of receiving instruction on 4 sessions, students took mini-tests. Their errors were isolated to become the focus of 4 teacher-guided feedback sessions. In the EI condition, teachers explicitly taught the mathematical material for all 8 sessions.

Results: Teacher time-on in the LFE condition produced a higher rate of learning than did teacher time-on in the EI condition. The learning benefit in the LFE condition was, however, inconsistent across teachers. Second-by-second analyses of classroom activities, directed at isolating learning-relevant differences in teaching style revealed that a highly interactive mode of engaging the students in understanding their errors was more conducive to learning than was teaching directed at getting to the correct solution, either by lecturing about corrections or by interaction focused on corrections.

Conclusion: These results indicate that engaging the students interactively to focus on errors, and the reasons for them, facilitates productive failure and learning from errors.

KEYWORDS
discovery learning, interactive feedback, learning from errors, productive failure, test effects

INTRODUCTION

The generation of errors, according to early learning theorists such as Bandura (1986) or Skinner (1953), was assumed to have adverse consequences for learning. It was thought that the commission of errors would entrench the errors themselves and that the correction of such errors was problematic. Empirical studies of the effects of error generation, however, have not substantiated these concerns. Indeed, studies have largely shown that, as long as corrective feedback is given, memory for the provided correct answer is usually *helped* rather than hurt by the subject's own prior error generation (Butterfield & Metcalfe, 2001, 2006; Clark & Bjork, 2014; Hays et al., 2013, Kornell et al., 2009; McDaniel et al., 2011; Metcalfe & Eich, 2019; Metcalfe & Finn, 2012; Metcalfe & Xu, 2018; Roediger & Finn, 2010; and see Metcalfe 2017 for a review). Here, such positive effects of errors as well as their potential use in fostering understanding are investigated.

Many studies have demonstrated beneficial learning effects from test taking. Indeed, research on the so-called 'testing effect' has led to considerable excitement about this method of enhancing learning (Brown et al., 2016; Kang et al., 2007; McDaniel et al., 2007; Pashler et al., 2005, 2007; Roediger & Karpicke, 2006; Roediger et al., 2010, 2011). The benefit to memory from retrieval practice of correct answers is firmly established (Agarwal et al., 2021; Karpicke, & Blunt, 2011; McDermott, 2021; Pan & Rickard, 2018). There is, however, another consequence of taking tests: people make errors.

The learning consequences of errors appear to depend upon how those errors are subsequently treated. Rather than necessarily being harmful, errors could, in principle, help learning. They could, for instance, set the stage for discovery learning (Bok, 2006; Kuhn et al., 2000; Metcalfe et al., 2023; Zhang & Fiorella, 2023 and see, Giebl et al., 2021; St. Hilaire et al., 2019) in which the learner is not provided with the target information or answer but must find it themselves. Unfortunately, a meta-analysis by Alfieri et al. (2011) indicated that discovery learning alone, with no feedback – or what they called 'unassisted' discovery learning – had limited success. Similarly, if a person makes a mistake on a test, and is not provided with corrective feedback, they rarely spontaneously provide the correct answer on retest (Pashler et al., 2005). Without feedback, and perhaps feedback of a particular sort (see, Zhang & Fiorella, 2023), the error is likely to persist.

A second meta-analysis by Alfieri et al. (2011), on 'guided,' 'enhanced' or 'enriched' discovery learning was more encouraging. When discovery learning was combined with subsequent feedback – by scaffolding, by the use of worked examples, or by elicited explanations – performance outshone that seen with explicit instruction. Work by Richland et al. (2009), suggests similar effects for errors. They showed that the time spent unsuccessfully trying to come up with an answer, while of little use if the answer is not subsequently found or provided, is positively related to learning once the correct answer is either discovered or given to the participants. Feedback, perhaps directed in a particular manner at student errors (see Freeman et al., 2014), appears to be crucial if those mistakes are to be converted into what is sometimes called '*productive* failure' (Kapur, 2008; Loibl et al., 2017; Loibl & Leuders, 2019; Zhang & Fiorella, 2023). Such feedback is the focus of the present study.

Knowledge of students' errors has a second potentially beneficial effect. It could position teachers to better appreciate and thereby counteract the misunderstandings of their students (see, Black & Wiliam, 1998; Rittle-Johnson, 2006; Wiliam, 2011). The errors students make on tests could potentially provide teachers with a roadmap for the content of the teaching needed to guide their students in overcoming their conceptual gaps.

Finally, the content (math, language learning, history, general information questions, etc.) of the test, and whether the test upon which the students' learning is measured is a 'test that counts' may matter. Clearly, while much can be learned from well-controlled laboratory experiments, it is possible that the effects produced under these highly controlled conditions may not play out in the real world (cf. McDaniel et al., 2007). It is equally conceivable – and perhaps even likely – that some of the benefits seen under lab conditions may be exacerbated rather than diminished when the stakes are high. In the study reported herein, the criterion test was the New York State Regents Common Core Algebra 1 examination – an examination that all public school students in New York State must pass in order to graduate from high school. It can provide the partial basis for students' later entrance into colleges of their choice, for the standing of the students' teachers, and for the ranking (and sometimes the fate) of their schools.

The present study focused on error-directed feedback, or what will be called learning from errors (LFE), as preparation for a test that counts (the Regents examination). Students already had a basic grounding in the to-be-learned Algebra 1 material from their regular classroom instruction. Here, in the LFE condition, a mini-testing session is followed the next day by a teaching/tutorial session directed at the errors the students had committed in the mini-test. This LFE method is contrasted with explicit instruction (EI). Both types of test-preparation were conducted by the same experienced teachers.

METHOD

Overview

Grade 8 students, who were required to pass the New York State Algebra 1 Regents examination for high school accreditation, participated in a 16-session, within-subjects, after-school tutorial program. In 8 of the sessions, students received EI. The other 8 sessions included 4 testing sessions in which questions from past Regents examinations were administered without instruction, followed by 4 LFE sessions. The study included cohorts in 2 years. In the first year, the criterion test was the Regents examination that the students sat for in June of 2016. The NYC Board of Education, provided access to the children's detailed, item-by-item data on their actual Regents examination performance. In the second year we used the same 2016 Regents examination, which was administered to the students in a separate session.

Participants

The participants were students in 8th Grade in a New York City public school who volunteered with consent from their parents. Students were not asked for individual demographic information, due to NYC Board of Education policy. The demographic breakdown of the school was 12% Asian, 23% Black, 37% Hispanic and 23% White; students with special needs: 12%; 51% male and 49% female, and 53% were below the poverty line. All students in Grade 8 who were expected to take the Regents examination in June of the year of the project, including special needs students, were invited to participate. Totally 177 students participated, with 175 completing most sessions and the pre-test and post-test. If a student missed a session, they continued to participate in the following sessions as soon as they could, and their data were included in the analyses. This study was approved both by the Columbia University Internal Review Board (protocol number AAAP7055) and by the NYC Board of Education.

Design

Students were randomly assigned to 4 teachers, except for a constraint imposed by the NYC School Board that no child be assigned to their own classroom teacher. This affected one teacher's assignment. The Regents examination is divided into questions that are officially designated as 'algebra' or 'function' questions, a designation we used to balance materials across teacher and condition (and a few problems on Statistics, Number and Quantity, that we ignored). The design was a 2 (Condition: EI, or LFE) × 2 (Materials: Algebra, or Functions) modified quasi mixed model within-subjects design. In the LFE condition, each test set consisted of 14 questions – 10 multiple choice questions and 4 constructed-response items. An additional question was included in one LFE session in year 1, but, because of time constraints, was eliminated in all other sessions. The materials were taught as a block of 8 sessions each. Algebra always was included in the first 8 sessions (because it was taught in class earlier in the year than Functions). There were 4 teachers, two of whom, in the first year, taught algebra in the EI condition and functions in the LFE condition, and 2 of whom taught algebra in the LFE condition and functions in the EI condition. In the second year, the teachers switched conditions so that, over the 2 years, all four teachers taught both algebra and functions in both EI and LFE condition.

Materials

A pre-test – the January 2016, Algebra 1 Common Core Regents examination – was administered 2 weeks prior to the test-prep sessions and provided baseline scores. The post-test was the June 2016 Algebra 1 Common Core Regents examination. Similar pre-test to post-test gains were observed for both cohorts. Results reported here are based on only algebra and the function questions.

The LFE test materials were questions – separated into algebra and functions – that had appeared on previous Algebra 1 Regents examinations (https://www.nysedregents.org). After all 16 sessions as well as the pre- and post-test sessions were completed, a questionnaire on growth mindset (https://mindsetonline.com/testyourmindset/ Dweck, 2008) was administered to participants. No differences were significantly correlated with this measure, so it will not be discussed further.

Procedure

In the EI condition, teachers were given a list of topics to cover each day by the head of mathematics at the school and created lesson plans and structured their teaching following their usual classroom practice. They used 'Let's Review: Algebra 1' (Rubenstein, 2015), in the Barron's series, as an authoritative resource that is specific to the Regents test, and could include problems for the children to work, at their own discretion under the guidance of the mathematics head. There were 8 such EI sessions.

In the LFE condition, the students took a mini-test on Day 1. The tests were scantron scored on the multiple choice questions, and hand-scored by a research assistant on the constructed answers. The errors were then computer tabulated by frequency and response, and a profile of their own students' errors was computer generated and provided to teachers for their Day 2 tutorial session. The students were given back their tests on Day 2, and the teachers were instructed by the experimenter to focus on the students' errors and to do whatever they deemed appropriate to ensure that the issues underlying the errors would not reoccur and that the students would learn from their errors. There were 4 mini-test sessions each of which was followed, the next day, by an LFE session.

All 12 teaching sessions (8 in the EI condition and 4 in the LFE condition) were audio–video recorded. In compliance with the Board of Education requirement, the video recordings were stored on computers with no connection to the internet, and all identifying markers of the students in the study (such as their faces or names) were scrubbed from the tapes before any analyses were conducted. Then the tapes were segmented into time units indicating differences in what was happening in the classroom on a second-to-second basis by one scorer. Then activity in each segment was coded as giving instructions, disciplining, distraction, teaching overall test-taking strategy, telling jokes, motivating the students, lecturing with a focus on the correct response, lecturing with a focus on examining the reason for a possible error, interactive teaching with a focus on getting to the correct answer, interactive teaching with a focus on the error, and miscellaneous, by two scorers. When the teacher was discussing a particular problem in the LFE condition, the problem number was coded along with the appropriate segment, allowing us to collate the problems (and the kind of teaching) with the errors that individual students had made on the tests. The scorers also coded whether the teaching style was interactive or lecture, and whether the teacher was discussing the error itself or was indicating a way to correctly solve the problem. These coding designations, made by two independent raters were nearly always the same, but in the rare cases of a disagreement the time was split across the conflicting designations.

RESULTS

The overall performance

The students' scores improved by 12.2% (SD = 9.8%) from pre-test to final Regents examination: pre-test (57.5%; SD = 17.4%); final examination (69.7%; SD = 17.4%), (t(174) = 16.5, p < .0001). As shown in Figure 1, whereas most students showed learning, i.e., post-test > pre-test, children who had scored lower on the pre-test exhibited more learning than did the students who had scored higher. A regression of students' learning scores on their pre-test scores indicated that pre-test scores were negatively associated with amount of learning (b = −.15, SE = .04, t(173) = −3.50, and p = .001).

To address the possibility that this differential improvement was due to a ceiling effect, we performed an analyses using proportion of possible improvement, computed as (post-test score − pre-test score)/(100 − pre-test score) for each participant. For example, if a student scored 80% correct on the pre-test, and 90% correct on the post-test, their corrected score was 10/20 or .50. If they got 60% on the pre-test, and 70% on the post-test their corrected score was 10/40 = .25. The regression indicated that students showed an improvement of about .20 of their own possible gain, and the effect of pre-test score was no longer significant (b = .17, SE = .15, t(173) = 1.18, and p = .241).

When simple overall scores were considered, there was an effect of type of materials, such that the pre-test–post-test score difference was greater for algebra (14.6%, SE = .86%) than that for function

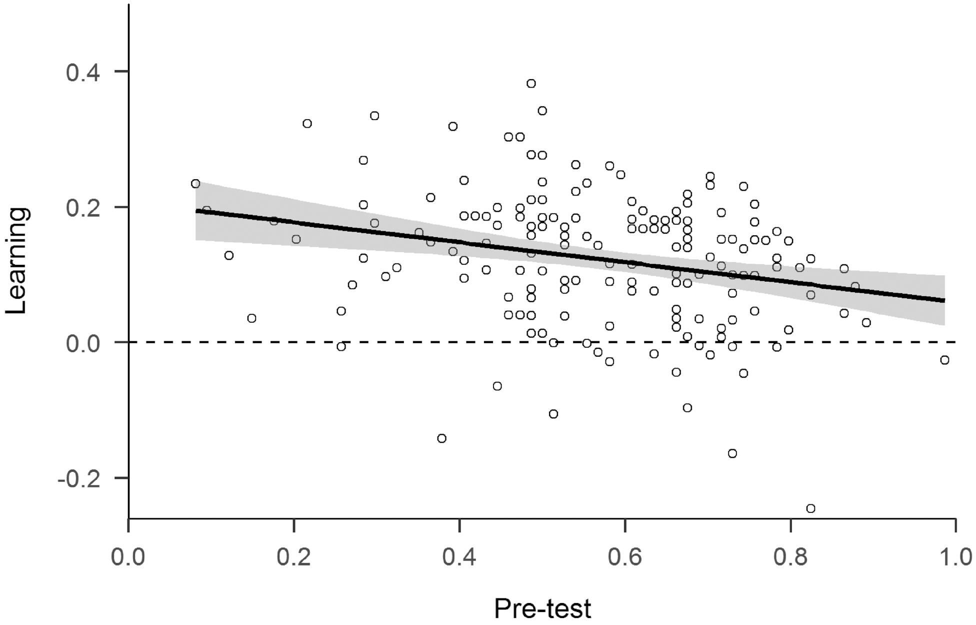

FIGURE 1 Learning as a function of students' scores on the pre-test. The regression line indicates the correlation between students' performance on the pre-test and their learning as indicated by the difference between pre- and post-test scores. Grey shading indicates 95% confidence interval.

questions (8.72%, SE = 1.11%). There was no difference between the EI (12.72%, SE = 1.06%) and the LFE (10.61%, SE = .97%) condition, *F*(1, 171) = 2.69, *p* = .103). Note, though, that the teachers taught only 4, 45 min sessions for a total of 180 min in the LFE condition (because there were 4 testing sessions in that condition during which the teachers did not teach), whereas they taught for all 8, 45 min sessions – for a total of 360 min – in the EI condition. Although there was no main effect of teacher (*F*(3, 171) = 1.29, *p* = .280), there was an interaction between condition and teacher, (*F*(1, 171) = 4.24, *p* = .006).

To further investigate, we computed 'teaching efficacy' – the learning scores *per hour* of teaching for each of the 4 teachers. The teaching efficacy was greater in the LFE condition (percentage gain per hour: 3.48%; *SE* = .32%) than in the EI (percentage gain per hour: 2.09%, SE = .18%; *F*(1, 171) = 17.94, *p* < .001). Again, there was an interaction between teacher and condition (*F*(3, 171) = 5.49, *p* = .001). As shown in Figure 2, all teachers elicited about the same learning gains from their students in the EI condition (*F*(3,171) = .89, *p* = .447). However, they varied greatly in the LFE condition (*F*(3,171) = 4.83, *p* = .003). In particular, in the LFE condition, the learning results of teachers 3 and 4 were very different, (*t*(82.98) = 3.85, *p* < .001). Whereas the effectiveness of teacher 4's teaching was about the same in the LFE and the EI conditions (*t*(42) = −.95, *p* = .347), teacher 3 showed much greater learning returns per hour of teaching in the LFE condition than in the EI condition (*t*(41) = 5.03, *p* < .001). The same was true for teacher 2 (*t*(44) = 2.93, *p* = .005). Teacher 1 numerically showed better returns in the LFE than those in the EI condition, but the difference was not significant (*t*(44) = 1.64, *p* = .108). It appears that the LFE method can be highly effective, but only for some teachers.

Error focus

One possible explanation for the learning differences shown in Figure 2 might have been that there were differences in the extent to which particular teachers actually focused on their students' errors.

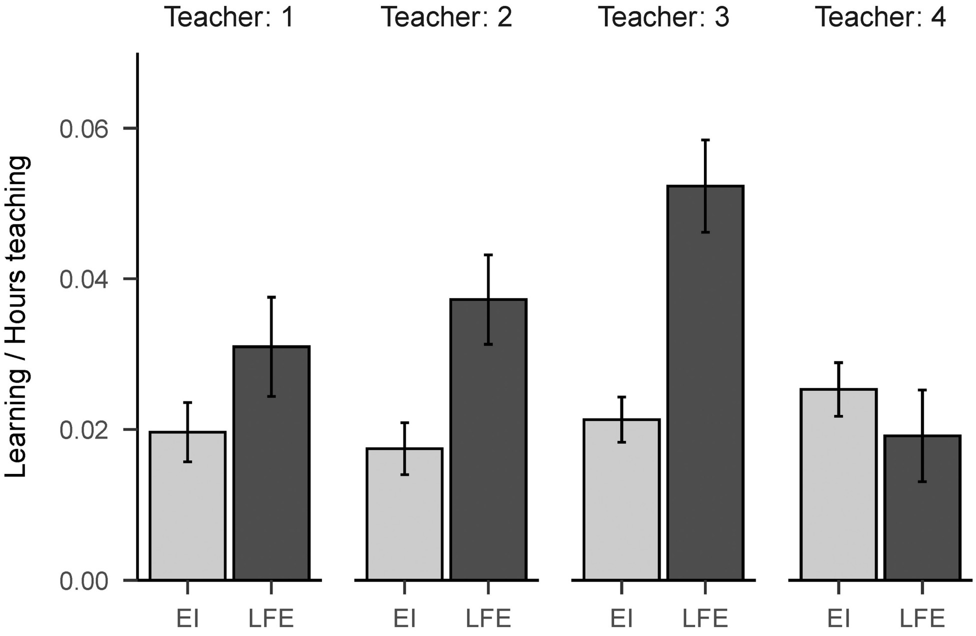

FIGURE 2 Mean learning per hour of teaching in the EI and LFE conditions by teacher. Error bars indicate standard errors.

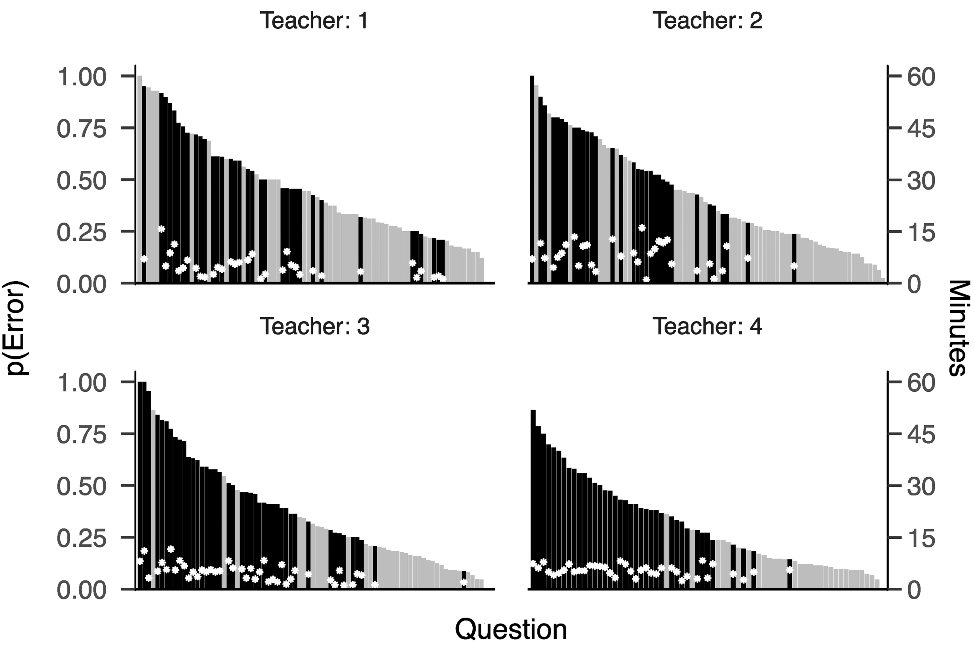

FIGURE 3 Profile of errors, and errors taught, for each teacher. The frequency of each error is given in order from left to right, with black bars indicating that the teacher taught that error in the feedback session, and grey bars indicating that that error was not discussed. White dots indicate the amount of time (noted on the right axis) that the teachers spent teaching each error. Notably, teachers 3 and 4 (the teachers who got the best and worst results) spent about the same amount of time per question.

Which problems each teacher chose to teach, and for how long, are shown in Figure 3 in which the frequency of each problem being an error made by students is rank ordered such that the most frequent errors are on the left and the least frequent errors are on the right. Whether the error was taught by the teacher is shown in black (if taught) and in grey (if not taught). Thus, if the teachers were teaching to most common errors, a great deal of black to the left hand side of the graph would be evident. Kruskal–Wallis gamma correlations (*G*) examining whether teachers addressed the questions on which most students erred were computed for each of the four teachers. All teachers taught to their students' errors ($G_{\text{teacher 1}} = .47$, $G_{\text{teacher 2}} = .69$, $G_{\text{teacher 3}} = .79$, $G_{\text{teacher 4}} = .93$). Although both teachers 3 and 4 taught to their students' errors quite consistently, Teacher 3 had the highest returns in terms of learning gains, whereas teacher 4 had the lowest.

We also quantified each teacher's tendency to teach to the students' errors by regressing whether the teacher taught a problem on the proportion of students who erred on that problem. The logistic regression included error proportion, teacher, and their interactions as predictors, and teacher 1 was treated as baseline. The error proportion predicted all teachers' probability to teach the question, but teachers 3 and 4 had stronger effects of error proportion on probability of teaching the question than did teacher 1 ($b = 5.67$, $SE = 2.32$, $z = 2.44$, $p = .015$; $b = 18.83$, $SE = 5.59$, $z = 3.37$, $p = .001$).

Finally, a signal detection framework was used in which *d*′ scores reflected the correspondence between the errors that each student made and the questions the teacher taught. For each student, a hit was designated when a student made a particular error and the teacher taught that item, and a false alarm was deemed to have occurred when the teacher taught an item that had not been an error for the particular student. Mean *d*′s for the students of each teacher were compared to zero. *t*-Tests indicated that all four teachers taught to the errors of their students ($M\ d'_{\text{teacher 1}} = .87$, $t(42) = 12.66$, $p < .001$; $M\ d'_{\text{teacher 2}} = 1.00$ $t(48) = 16.38$, $p < .001$; $M\ d'_{\text{teacher 3}} = 1.14$, $t(47) = 15.63$, $p < .001$; $M\ d'_{\text{teacher 4}} = 1.14$, $t(43) = 18.85$, and $p < .001$).

Teaching to errors and learning

The relation between the extent to which the teacher taught to the errors (as given by the *d*'s) and the individual students' learning is represented in Figure 4. A regression of students' learning scores on their d's, indicated that *d*' was not statistically significant ($b = -.007$, $SE = .022$, $t(169) = -.32$, $p = .748$). The interaction when teacher was added as a predictor was also not significant ($F(3, 163) = 2.47$, $p = .064$). It appears, from these results, that all of the teachers taught to the errors of their students, and that the extent to which they did so did not predict student learning.

Style of teaching

The videotapes of each session were analysed on a second-by-second basis to determine how much time was spent in various classroom activities. Time periods in a particular activity were only included in this analysis if the activity lasted for at least 4 s. The categories of activities that were coded were (a) disciplining the students, (b) distraction, (c) group work, (d) individual work, (e) organizational instruction, such as telling the students to take out their books, go to certain pages, etc., (f) jokes, (g) motivational pep talks, including encouragements, reassurance, etc., (h) test-taking strategy, such as telling the students to fill in all questions, to use an educated guess if they didn't know the answer, or to use their graphing calculators to check answers, etc., (i) teaching in an interactive manner, in which the teacher, for instance, asked pointed questions and allowed the students guide the discussion and (j) teaching by lecturing, in which the teacher spoke with little to no input from the students.

As shown in Figure 5, all teachers spent nearly all of their time teaching (as shown by columns I and J); there was little time spent on other activities. In addition, 'individual work' (column d) resulted selectively in the EI condition because all teachers sometimes had students work, on their own, on problems

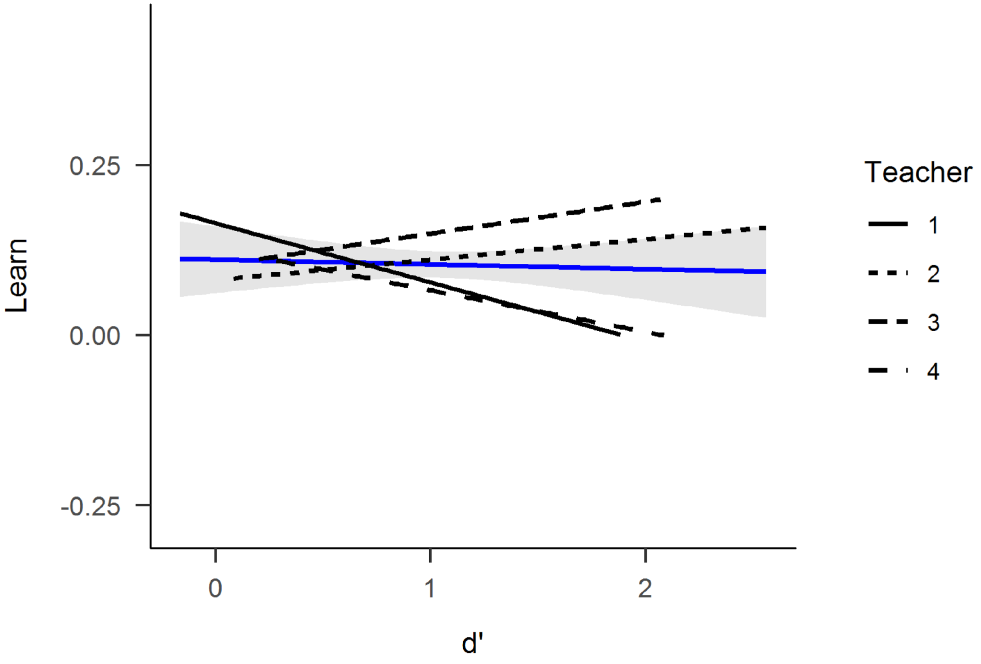

FIGURE 4 The relation, in the LFE condition, between student learning and the directedness of the teaching to the student's own errors, as indicated by the *d*' relating each individual's errors to the questions taught during the feedback sessions, for each of the 4 teachers. The blue line is the overall trend. Grey shades are 95% CIs.

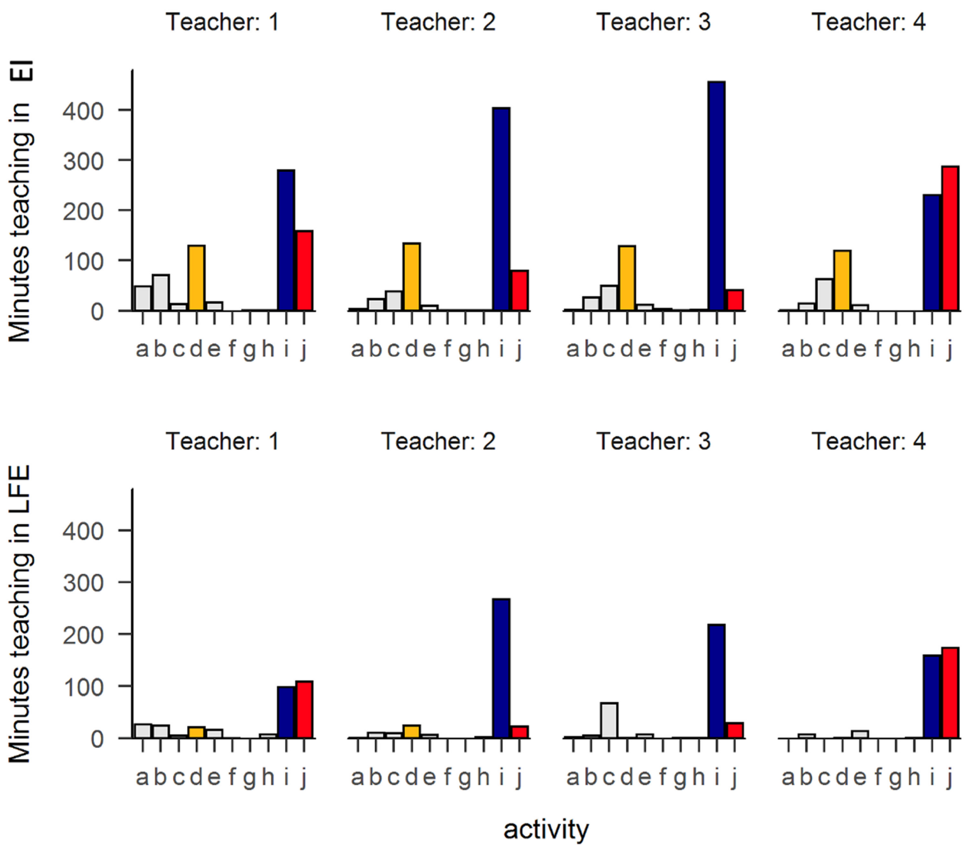

FIGURE 5 Time each teacher allocated to various activities in the EI (top panel) and LFE (bottom panel) condition. The activities were (a) discipline, (b) distraction, (c) group work, (d) individual work, (e) organizational, (f) jokes, (g) motivational pep talk, (h) test taking strategy, (i) teaching–interactive and (j) teaching–lecture.

they had just reviewed. It is also apparent, even from this initial survey, that teachers 2 and 3 spent their time predominantly in interactive teaching and spent little time lecturing, whereas teacher 4 and to a lesser extent teacher 1, spent more time lecturing.

Teaching mode and focus

To drill down more on *how* the teachers taught, we isolated those cases during which the teachers were teaching to errors in the LFE sessions, and sorted into mode (lecturing or interactive) and focus (correction or error). 'Correction' meant that the teacher was dwelling upon how to solve the problem correctly. 'Error' meant that the teacher was delving into the nature of the errors – why the students had made them, what the difficulty in the logic was, and/or how to recognize and circumvent such mistakes in the future.

The proportions of teaching time within LFE on these four kinds of teaching (lecturing/correction; lecturing/error-focused; interactive/correction; interactive/error-focused), for each of the 4 teachers, are illustrated in Figure 6. Teachers 1 and especially teacher 4 spent most of their time focused on how to get the correct answer rather than exploring the errors. Teacher 4, in particular, spent much time lecturing about how to get the correct answer. Teachers 2 and 3 focused more on discussing the errors, spending about an equal amount of time discussing what had gone wrong and how to generate the correct answer. Both teachers did so in a highly interactive manner and spent little time lecturing.

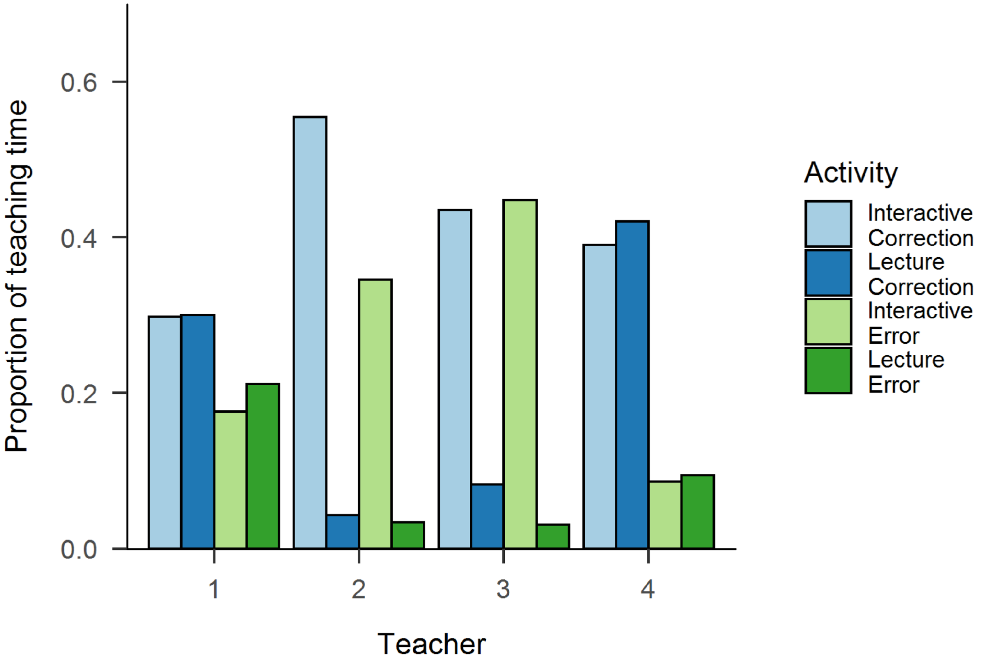

FIGURE 6 Conditionalized proportion of time in the LFE condition spent in interactive–correction, lecture–correction, interactive–error and lecture–error teaching. Intervals in which it was not clear whether the teaching was directed at the correct answer or the error were not included.

To compare the teachers to one another statistically, we computed the proportion of teaching time per question in either interactive/correction, interactive/error, lecture/correction, lecture/error, and conducted a 2 (Mode: Interactive vs. Lecture) × 2 (Focus: Error v. Correction) × 4 (teacher) ANOVA. Teachers spent more time in interactive teaching ($M = .34$, $SE = .01$) than lecturing ($M = .16$, $SE = .01$; $F(1,580) = 113.38$, $p < .001$). They also spent more time teaching how to get the answer correct ($M = .30$, $SE = .01$) than in discussing the errors ($M = .20$, $SE = .01$; $F(1,580) = 35.97$, $p < .001$).

Because we conditionalized total teaching time to be 100% of each teacher's time, there was no main effect of Teacher. However, there were interactions between Teacher and Mode (Interactive or Lecture), $F(3,580) = 55.71$, $p < .001$, and between Teacher and Focus (Error or Correction), $F(3,580) = 24.25$, $p < .001$. There was no interaction between Mode and Focus, $F(1,580) = .41$, $p = .52$. There was, though, a 3-way interaction among mode, focus and teacher, $F(3,580) = 4.74$, $p = .003$. As shown in Figure 6, Teacher 1 spent more time on corrections than discussing the errors and was evenly split between interactive and lecture style. Teacher 4 was even more strictly focused on correction, and rarely talked about the nature of the errors. He was also about evenly split between lecture and interactive modes. Teachers 2 and 3 showed a different pattern. They spent most of their time teaching interactively regardless of whether it was examining an error or the correct procedure. These teachers also spent a much higher proportion of their time exploring the errors, nearly always interactively.

The relation of teaching style to student learning

We investigated how the amount of time spent teaching in interactive–correction, lecture–correction, interactive–error and lecture–error related to individual students' learning. To do so, we isolated the problems on which each student had made errors, and tabulated the amount of time spent on their particular errors in each of the four modes. (To have enough data for the analysis we collapsed over teacher.) This individualized teaching-time score was used to predict learning in each of the four modes. We conducted 4 separate regressions, one for each of the teaching modes – interactive–correction, lecture–correction, interactive–error and lecture–error.

As is shown in Figure 7, time spent on a student's error in the interactive–correction mode did not affect learning ($b = -.02$, $p = .688$). Time that teachers spent on students' errors in both the lecture–correction mode ($b = -.16$, $p < .001$) and in the lecture–error mode ($b = -.32$, $p = .0037$) *negatively* impacted learning (and see Knight & Wood, 2005). Only in the interactive error mode did more time spent focusing on the error have a positive effect on learning ($b = .17$, $p < .001$). The gain of about a 17% increase in learning per hour of such teaching was sizable.

DISCUSSION

Despite the fact that all teachers, in the LFE condition, addressed the problems on which the students had given erroneous responses, not all teachers stimulated equal learning benefits. Simply providing correct feedback about errors did not appear to be enough. Of course, if the teachers had spent their teaching time on questions that all of their students had already answered correctly, we would not

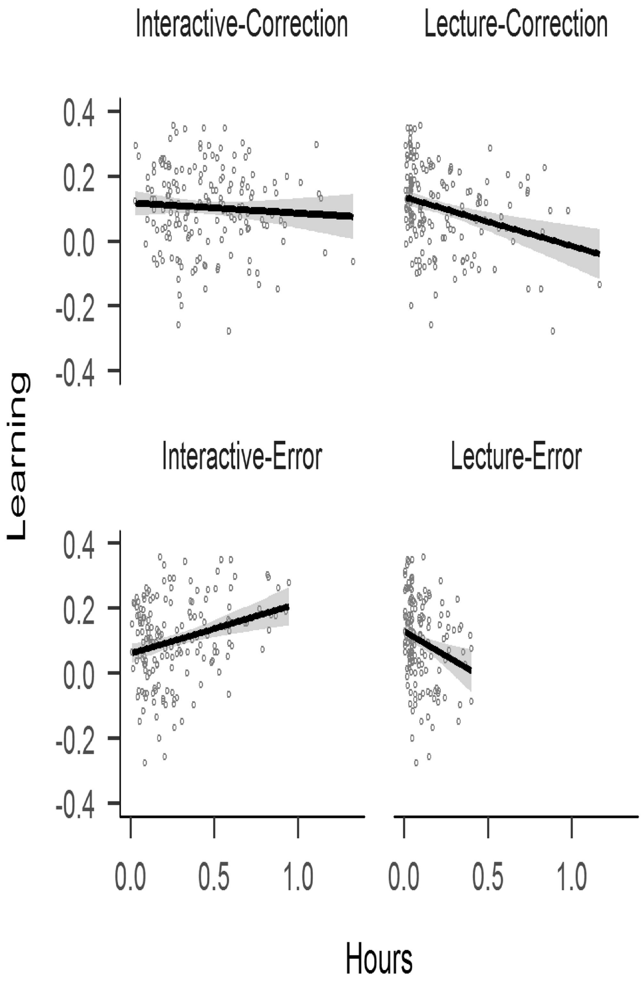

FIGURE 7 Regressions of teaching mode against learning. Grey shades indicate 95% CIs.

expect the children to learn much. We cannot say what would have happened had teachers taught in this way, since none of our teachers did it. Teaching to the errors is undoubtedly important. But it seems to be only the first step.

How the students are taught appears to be essential. Interacting with the students on understanding what the errors are and why they occurred, and collaborating on ferreting out why mistakes were made and how to avert them in future seems to be the most effective approach (see, e.g., Fiorella & Mayer, 2016; Frazier et al., 2021; Rittle-Johnson, 2006; Rittle-Johnson et al., 2017; Weiman et al., 2008). Of course, an important limitation in our study – and one that deserves future investigation – is that mode of teaching was not manipulated. Instead, it was merely observed and correlated with learning. It was also, though, the preferred mode of teaching used by the two teachers who got the best results in the LFE condition. Perhaps, though, it was simply that these two teachers were better teachers, and the correlation (driven by them) was spurious.

There are two indications in our data that argued against this interpretation. First, all four teachers achieved the same learning in the EI condition. If two of the teachers were simply better instructors, they should have been better under EI conditions as well. Second, the two teachers who returned the best learning results also spent considerable time not in the Interactive Error mode but in the interactive correction mode. If these two teachers were simply the best teachers, and the benefits were not due to teaching selectively in the interactive error mode, then one would expect time spent in this corrective mode to foster learning as well. But it did not.

The non-collaborative (lecture) approach was not related to high learning in our study; nor was ignoring the reasons for the errors and focusing only on their correction productive. Only collaboratively and interactively engaging with the reasons for the errors helped. So, it is plausible that there may be a special kind of teaching that is particularly conducive to evoking learning gains: highly supportive, interactive, exploration of the errors that students made, including discussion *with* them – rather than *at* them – about how to overcome those errors.

It is of some interest that the teaching modes in our study corresponded fairly closely to those that Stigler et al. (1999; Stevenson & Stigler, 1994; Stigler & Hiebert, 2009) observed when they investigated teaching methods in Japanese and American classrooms, in the TIMMS study. That study investigated the methods of teaching that contributed to the high math scores of Japanese as compared to American students. In that study, American teachers appeared to be error avoidant – ignoring students who made errors and pivoting to those whom they knew would answer correctly. In our study, a teaching strategy in which a strong emphasis was placed only on the correct answer was also not effective: lecture–correction was harmful, and interactive–correction had no effect. The (effective) Japanese teachers in the TIMMS study, engaged with errors, discussing and exploring them with their students. This mode of interactive teaching resonates with the style in our study – interactive–error – that produced the strongest learning gains.

While our results are suggestive, there are limitations to our study. We mention just a few. First, the study used only four teachers. A much larger sample is needed. Second, the student population was highly motivated. The generality to less motivated students needs to be tested. Most importantly, the analyses we conducted on teaching style – which comprise a core contribution of this article – were exploratory rather than confirmatory, a kind of qualitative study with numbers. Before any firm conclusions can be drawn, further investigations are needed in which teaching style, in interaction with errors, is manipulated.

The findings of this study may help to shed light on an important debate within education. For many authors, the purpose of formative assessment is primarily to inform teachers about their students' weaknesses so that the teacher may make instructional adjustments. However, merely identifying errors, and teaching students how to correct them, may not be the most effective strategy. Wiliam and Thompson (2008) emphasize that formative assessment is something that teachers do *with* students rather than *to* them. Much more work needs to be done to identify the mechanisms at work here, but it seems that it is the combination of interactive discussion and the focus on errors that is important. Without the focus on errors, the discussion may be pitched at a level that is outside the student's own region of proximal learning (Xu & Metcalfe, 2016).

Without the interaction and self-involvement, the task may be less interesting and motivating because of the lack of personalization. Thus, both the focus on errors and interactive exploration of the reasons for those errors may be needed to produce the degree of engagement needed to create '*desirable* difficulties' for the students (Bjork, 2017; Bjork & Bjork, 2014; Metcalfe, 2011).

AUTHOR CONTRIBUTIONS

Janet Metcalfe: Conceptualization; investigation; funding acquisition; writing – original draft; methodology; validation; visualization; writing – review and editing; project administration; data curation; supervision; resources. **Judy Xu:** Conceptualization; investigation; methodology; validation; visualization; writing – review and editing; formal analysis; data curation; supervision; project administration. **Matti Vuorre:** Conceptualization; investigation; methodology; validation; visualization; writing – review and editing; formal analysis; data curation; supervision; project administration. **Robert Siegler:** Conceptualization; methodology; validation; visualization; writing – review and editing; supervision; investigation. **Dylan Wiliam:** Conceptualization; investigation; methodology; validation; visualization; writing – review and editing; supervision. **Robert A. Bjork:** Conceptualization; investigation; methodology; validation; visualization; writing – review and editing; supervision.

ACKNOWLEDGEMENTS

We express our special thanks to Miriam Nightengale, Laura Schultz, Paul A. Bloom, Michael Helzer, Thomas Wilson, Roy Prieb and his team, and to the dedicated teachers and students, who contributed so much to this project.

FUNDING INFORMATION

This research was funded by United States Department of Education Institute of Education Science grant R305A150467 to the first author. The authors are solely responsible for the data and interpretations herein.

CONFLICT OF INTEREST STATEMENT

All authors declare that they have no known conflicts of interest.

DATA AVAILABILITY STATEMENT

The data supporting these findings are available upon request. Our IRB clearance--which also involves clearance from the NYC School Board-- has imposed restrictions governing the handling of the data of the children. Thus, please contact the first author at jm348@columbia.edu for access to de-identified data.

ETHICS STATEMENT

This research, which involved human participants, was approved by the Columbia University Internal Review Board under Protocol Number: AAAP7055. All children participants and their parents/guardians signed informed assent/consent. All adult participants in the research signed informed consent forms. The research reported here is original and the manuscript is not under submission to any other journal.

ORCID

Janet Metcalfe https://orcid.org/0000-0003-3286-9475
Matti Vuorre https://orcid.org/0000-0001-5052-066X

REFERENCES

Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*(4), 1409–1453.

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1–18.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bjork, E. L., & Bjork, R. A. (2014). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher & J. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (2nd ed., pp. 59–68). Worth.

Bjork, R. A. (2017). Creating desirable difficulties to enhance learning. In I. Wallace & L. Kirkman (Eds.), *Best of the best: Progress* (pp. 81–85). Crown House Publishing.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*, 7–74.

Bok, D. (2006). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton University Press.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2016). *Make it stick: The science of successful learning*. Harvard.

Butterfield, B., & Metcalfe, J. (2001). The high confidence error hyper-correction effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1491–1494.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*, 69–84.

Clark, C. M., & Bjork, R. A. (2014). When and why introducing difficulties and errors can enhance instruction. In V. A. Benassi, C. E. Overson, & C. H. Hakala (Eds.), *Applying science of learning in education: Infusing psychological science into the curriculum* (pp. 20–30). Society for the Teaching of Psychology. http://teachpsych.org/ebooks/asle2014/index.php

Dweck, C. S. (2008). *Mindset: The new psychology of success*. Random House Digital.

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717–741.

Frazier, L. D., Schwartz, B. L., & Metcalfe, J. (2021). The MAPS model of self-regulation: Integrating metacognition, agency, and possible selves. *Metacognition and Learning*, *16*, 297–318.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 8410–8415.

Giebl, S., Mena, S., Storm, B. C., Bjork, E. L., & Bjork, R. A. (2021). Answer first or Google first? Using the internet in ways that enhance, not impair one's subsequent retention of needed information. *Psychology Learning and Teaching*, *20*(1), 58–75.

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(1), 290–296.

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

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26*(3), 379–424.

Knight, J. K., & Wood, W. B. (2005). Teaching more by lecturing less. *Cell Biology Education*, *4*, 298–310.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*, 989–998.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction*, *18*(4), 495–523.

Loibl, K., & Leuders, T. (2019). How to make failure productive: Fostering learning from errors through elaboration prompts. *Learning and Instruction*, *62*, 1–10.

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychological Review*, *29*, 693–715.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*, 399–414.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*, 200–206.

McDermott, K. D. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology*, *72*, 609–633.

Metcalfe, J. (2011). Desirable difficulties and studying in the region of proximal learning. In *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 259–276). Psychology Press.

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465–489.

Metcalfe, J., & Eich, T. S. (2019). Memory and truth: Correcting errors with true feedback versus overwriting correct answers with errors. *Cognitive Research: Principles and Implications*, *4*, 1–18.

Metcalfe, J., & Finn, B. (2012). Hypercorrection of high confidence errors in children. *Learning and Instruction*, *22*(4), 253–261.

Metcalfe, J., Vuorre, M., Towner, E., & Eich, T. S. (2023). Curiosity: The effects of feedback and confidence on the desire to know. *Journal of Experimental Psychology: General*, *152*(2), 464–482.

Metcalfe, J., & Xu, J. (2018). Learning from one's own errors and those of others. *Psychonomic Bulletin & Review*, *25*(1), 402–408.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756.

Pashler, H., Bain, P., Bottge, B., Graesser, A., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning: A practice guide*. National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/ncee/wwc/practiceguides/

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.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*(1), 1–15.

Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM Mathematics Education*, *49*, 599–611.

Roediger, H. L., Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *New frontiers in applied memory* (pp. 13–49). Psychology Press.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*, 382–395.

Roediger, H. L., & Finn, B. (2010). The pluses of getting it wrong. *Scientific American Mind*, *21*(1), 38–41.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.

Rubenstein, G. M. (2015). *Let's review: Algebra 1*. Simon & Schuster.

Skinner, B. F. (1953). *Science and human behavior*. MacMillan.

St. Hilaire, K. J., Carpenter, S. K., & Jennings, J. M. (2019). Using prequestions to enhance learning from reading passages: The roles of question type and structure building ability. *Memory*, *27*(9), 1204–1213.

Stevenson, H., & Stigler, J. W. (1994). *The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education*. Simon & Schuster.

Stigler, J. (1999). *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States*. U.S. Dept. of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Stigler, J. W., & Hiebert, J. (2009). Closing the teaching gap. *Kappan*, *91*, 32–37.

Weiman, C. E., Perkins, K. K., & Adams, W. (2008). Oersted Medal Lecture 2007: Interactive simulations for teaching physics: What works, what doesn't, and why. *American Journal of Physics*, *76*(4), 393–399.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*, 3–14.

Wiliam, D., & Thompson, M. (2008). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Routledge.

Xu, J., & Metcalfe, J. (2016). Studying in the region of proximal learning reduces mind wandering. *Memory & Cognition*, *44*(5), 681–695.

Zhang, Q., & Fiorella, L. (2023). An integrated model of learning from errors. *Educational Psychologist*, *58*(1), 18–34.

How to cite this article: Metcalfe, J., Xu, J., Vuorre, M., Siegler, R., Wiliam, D., & Bjork, R. A. (2024). Learning from errors versus explicit instruction in preparation for a test that counts. *British Journal of Educational Psychology*, *00*, 1–15. https://doi.org/10.1111/bjep.12651