American Educational Research Journal
Fall 2003, Vol. 40, No. 3, pp. 769–801

Effect of Paced and Unpaced Practice on Skill Application and Retention: How Much Is Enough?

Normand Péladeau
Provalis Research
Jacques Forget and Françoys Gagné
Université du Québec à Montréal

*This study examined the relative benefits of mastery learning, overlearning, and fluency-building instructions for academic performance and long-term retention. College students enrolled in introductory quantitative methods classes (*n *= 168) were asked to practice every week with a computerized flash-card program until they attained various mastery criteria. The results confirmed that practicing until mastery improved individual exam scores, group success rates, and long-term retention. Moreover, overlearning provided additional benefits, especially in long-term retention. However, fluency-building instructions did not further increase academic achievement or long-term retention. Despite the alleged detrimental effects of drill and practice on motivation, a positive relationship was found between amount of practice and attitudes toward the course, the subject matter, and practice activities.*

Keywords: academic achievement, academic performance, fluency, long-term retention, mastery learning, overlearning.

Much research shows that the amount of time spent by students on academic tasks in the classroom correlates highly with academic achievement (Fischer, Berliner, Filby, Marlianve, Cahen, & Dishaw, 1980; Stallings, 1975). However, according to Stallings (1980), simply increasing time-on-task does not guarantee improved learning or better academic performance; the nature of the activities involved must also be taken into consideration. Using a fine-grained analysis of classroom activities, Greenwood, Delquadri, and Hall (1984) found that some activities contribute more than others to the acquisition of the target skills and to the academic success of students. They proposed, as an alternative to the traditional time-on-task notion, the concept of "opportunity to respond," which can be defined as the occurrence of teacher-formulated instruction (e.g., questions asked, materials presented, prompts) followed by appropriate academic responding (Greenwood et al., 1984). They found that the frequency of opportunities to respond predicted academic performance better than time-on-task measures. Thus any instructional antecedents and methods that increase the rate of correct academic responses for most students constitute efficient means for increasing academic achievement (Greenwood et al., 1984; Greenwood, Hart, Walker, & Risley, 1994). These findings confirm earlier observations made by Rosenshine and Berliner (1978), namely, that the most effective teaching methods are characterized by a pattern of controlled practice, consisting of factual questions, student academic responding, and adult academic feedback. Heward (1994) suggested that the frequency count of students' responses to the curriculum within a given instruction period may well be the most useful and readily available measure to assess how active a teaching method is. It can be argued that the capacity to produce a high "response frequency" precisely explains the well-documented effectiveness of computer-assisted instruction (Gillingham & Guthrie, 1987; Kulik & Kulik, 1987), Direct Instruction (Adams & Engelmann, 1996; Stallings, 1975; White, 1988), peer-tutoring (Cohen, Kulik, & Kulik, 1982; Delquadri, Greenwood, Stretton, & Hall, 1983), and teaching techniques such as choral responding, flashcards, practice sheets, and guided notes (Heward, 1994).

Although there is ample empirical evidence substantiating the relationship between response opportunities and academic performance, much remains to be discovered about which conditions best foster optimal learning performances. Many recurring questions concern the appropriate amount of practice (Ebbinghaus, 1885/1964; Heward, 1994; Judd & Glaser, 1969; Rosenshine & Berliner, 1978; Thorndike, 1921). How many times should students practice a specific skill? When should we interrupt practice on a specific topic and move on to something else? How should we schedule practice periods to optimize learning without submitting students to useless practice that can produce negative attitudes toward such a repetitive task? The most widely recognized criterion comes from the various instruction models grouped under the "mastery learning" perspective, such as Bloom's Learning for Mastery (Block & Anderson, 1975; Bloom, 1976) and Keller's Personalized System of Instruction (Keller, 1968). In these models, the progression of students through the learning units is contingent on attaining a mastery criterion,

Normand Péladeau is a Research Consultant and Software Designer at Provalis Research, 2414 Bennett Avenue, Montreal, QC, Canada H1V 3S4; e-mail *peladeau@simstat.com*. His areas of specialization are instructional design and evaluation, research methodology, and content analysis.

Jacques Forget is Professor of Psychology at Université du Québec à Montréal, P.O. Box 8888, Downtown Station, Montreal, QC, Canada H3C 3P8; e-mail *forget.jacques@uqam.ca*. His areas of specialization are applied behavior analysis, childhood disorders, Precision Teaching, and history of psychology.

Françoys Gagné is an Honorary Professor (retired), Department of Psychology, Université du Québec à Montréal, P.O. Box 8888, Downtown Station, Montreal, QC, Canada H3C 3P8; e-mail *gagne.francoys@uqam.ca*. His areas of specialization are gifted education, psychometrics, and talent development.

most often defined as an accuracy level of 80% to 90% correct responses. Students who are unable to reach that level are provided with corrective activities, such as alternative reading or peer tutoring. They then take the test again, repeating the sequence until the mastery criterion is attained. The effectiveness of these methods, when compared with traditional teaching methods, has been clearly demonstrated (Kulik, Kulik, & Bangert-Drowns, 1990; Kulik, Kulik, & Cohen, 1979; Guskey & Gates, 1986; Slavin, 1987). At the same time, some empirical evidence leaves open the question whether such a single accuracy criterion suffices to define true mastery.

Overlearning Research

One such area of study is the overlearning effect. Research findings on overlearning show that continuing practice after learners have attained a high-accuracy criterion may improve retention (Driskell, Willis, & Cooper, 1992). Many educational researchers have also stressed the importance of engaging students in frequent revisions of mastered skills (Brophy & Good, 1986; Hirsch, 1996; Rosenshine & Stevens, 1986; Vockell & Schwartz, 1988). Despite a large body of data supporting the positive effect of overlearning on retention, and despite the wide acceptance of the value of overlearning in education, research on this topic suffers from several methodological shortcomings that justify some concerns about its practical value for education. Some researchers have stressed the low ecological validity of many studies on memory (Bahrick, 1979; Neisser, 1983; Semb & Ellis, 1994). According to Bahrick, memory research has contributed little to our understanding of the acquisition and retention of complex knowledge under ecologically realistic conditions. Studies on overlearning raise similar objections. For example, most such studies have been performed in laboratory settings or have involved trivial or circumscribed learning tasks (e.g., lists of nonsense syllables, paired associates, spelling of short lists of words). Very few studies have been conducted in classroom settings using meaningful academic content, and they have generally yielded inconclusive results. For example, Reynolds and Glaser (1964) taught 75 high school students to recognize science terms by using various levels of learning; no difference in retention was observed after 2 days or 3 weeks. That study exemplifies another common shortcoming of studies on overlearning: Most of them use very short retention periods. For example, in the meta-analysis by Driskell et al. (1992), the median retention interval for comparisons involving cognitive tasks was 2 days, and the longest was 28 days. Both the lack of implementation in applied settings and the short retention measures are especially troublesome if we consider the conclusion by Driskell et al. that for cognitive tasks the increased retention due to overlearning tends to dissipate after only 5 to 6 weeks, a time span few educators would consider of any practical value. Although some correlational studies of teacher behavior suggest that engaging students in overlearning produces higher academic achievement (Brophy & Evertson, 1976; Brophy & Good, 1986) or that effective teaching methods often provide periodic reviews of learned material (Rosenshine & Stevens, 1986), these studies offer only indirect proof for the effectiveness of overlearning in school settings. Furthermore, they do not permit quantification of the benefits gained from repeated practice, nor do they offer any guidelines for the amount of overlearning needed to ensure long-term retention.

Precision Teaching and Fluency Research

Practitioners and scholars associated with the Precision Teaching movement (Binder, 1993) have also questioned the appropriateness of operationally defining a mastery criterion solely in terms of accuracy and have tried to provide answers to the "how much practice is needed" question. According to them, students who want to progress and become competent on a given composite skill or knowledge task must achieve both accuracy and speed on its components and prerequisites (Binder, 1993, 1996; Kubina & Morrison, 2000). For this reason, Precision Teaching specialists propose that progress through curriculum hierarchies should not be based on accuracy-only criteria but, rather, on fluency criteria, fluency being defined as accuracy plus speed (Binder, 1996). Several experimental and applied studies in this area confirm that practicing beyond a traditional mastery criterion up to a high fluency level significantly improves not only the students' academic performance and retention (Ivarie, 1986; Olander, Collins, McArthur, Watts, & McDade, 1986) but also their endurance, defined as the maintenance over time of highly accurate and fast responses (Binder, 1996; Binder, Haughton, & Van Eyk, 1990). Moreover, fluency facilitates the transfer of these newly acquired skills to applied settings or to more complex skills (Bucklin, Dickinson, & Brethower, 2000; Evans & Evans, 1985; Haughton, 1972; Johnson & Layng, 1992).

However, because (a) fluency training always involves practicing beyond initial mastery, and (b) repeated practice is accompanied by a decrease in response latency (Judd & Glaser, 1969; Newell & Rosebloom, 1981), it becomes difficult to ascertain whether the observed benefits of fluency training are caused by the achieved fluency level or simply by the amount of additional practice needed to establish it. This situation has led Dougherty and Johnson (1996) to suggest that fluency might not be distinguishable from overlearning or automaticity. However, Binder (1996) argues that overlearning could be considered the procedural description of the phenomenon, whereas fluency could be viewed as its measurable outcome. In this respect, fluency could nevertheless be useful, both as a measure of the degree of learning that occurs beyond a 100% accuracy level and as a criterion to decide when practice should be stopped. Judd and Glaser (1969) made a similar proposal in their studies of the effect of overlearning on response latency. They argued that response latency could serve as an accurate measure of learning progress and could thus be used for instructional decision making, especially in situations where response accuracy was high.

The emphasis that Precision Teaching places on measuring performance through response frequency and using high-paced prompting and fluencycoaching techniques to encourage students to engage in high-speed responding has led some researchers and teachers to make an even stronger claim (Berquam, 1981; Brown, Dunne, & Cooper, 1996; Kelly, 1996; McCarty, 1999). According to them, if one recognizes that response fluency accurately measures the degree of learning, it follows that (a) students can be specifically instructed to increase that frequency beyond what could be obtained by simply engaging them in overlearning trials, and (b) such an enhanced fluency provides benefits beyond the incidental fluency naturally ensuing from repeated practice. This hypothesis implies that response frequency would not be just an epiphenomenon of overlearning practice but could be controlled by antecedent stimuli such as pacing instructions or incentives. There is no doubt that such a demonstration would have great educational value: It would represent, much like the well-known spacing effect (Dempster, 1988), a way to obtain greater learning without additional practice. To our knowledge, no study has investigated whether subjects can specifically learn to decrease their response latency by any means other than repeated practice. There is, however, some empirical evidence suggesting that paced practice can provide additional benefits over unpaced overlearning practice. For example, in an experimental study of the performance of two groups of third graders on a paired-associates task, Berquam concluded that fluency-trained students had higher retention levels than students in an unpaced practice condition, even though the latter group completed 60% more practice problems. Unfortunately, the author did not report any accuracy measures, using only response rates as indicators of retention. Also, although he found differences on three of his four measures of retention, he failed to demonstrate any impact on the most relevant one. Moreover, he did not adequately control for prior differences in performance. Shirley and Pennypacker (1994) used a single-case experimental design to compare the effects of various fluency criteria on the spelling performance of two eighth graders; they observed higher retention rates from speeded practice for one student but not for the other. On a spelling instruction task, Ormrod and Spivey (1990) measured retention after 3 weeks; they found no statistically significant difference in accuracy between paced and unpaced overlearning practices. Nevertheless, their data showed a difference of 0.36 standard deviation in favor of the paced practice group, suggesting that statistical significance might have been achieved with a larger sample.

In addition to insufficient controls for amount of practice, many fluency studies suffer from inadequate controls for initial differences in student abilities. Most of these studies are correlational in nature. For instance, Binder (1984, quoted in Binder, 1996) found that students who achieved higher response rates showed smaller performance decrements when asked to maintain their performance level for long intervals. However, because shorter response latency is a well-known characteristic of high-ability students (McFarland, 1928, 1930; Spearman, 1914; Tinker, 1931; Wolf & Stroud, 1961), it is not clear whether the benefits observed for students who achieved high fluency levels result from the practice activity itself or from preexisting high intellectual aptitudes.

In summary, although many applied studies in educational settings have confirmed the importance of attaining high accuracy levels, the importance of engaging students in overlearning activities beyond that initial mastery level remains to be demonstrated. Much also needs to be learned about how fluency-building instructions in the form of paced practice affect the development, maintenance, and application of academic knowledge and skills. Finally, irrespective of any particular practice instruction, it is not clear whether attaining high fluency produces any specific academic benefits.

Purpose of the Study

We designed the present study to investigate the relative benefits of mastery learning and of paced and unpaced overlearning for academic performance and long-term retention. We tried to circumvent limitations of previous studies by (a) controlling for initial differences in cognitive abilities, (b) controlling for the amount of practice, and (c) experimenting in natural academic settings with meaningful academic content. We posed three main research questions:

1. Does attaining mastery improve academic performance and retention?
2. Does overlearning lead to better academic performance and long-term retention than terminating practice when initial mastery is attained?
3. Does paced practice, that is, instructing students to increase their correct-response rate per minute, provide any additional benefits beyond what can be obtained by mere additional practice?

The study was also designed to evaluate students' attitudes toward the subject matter and toward computerized practice under the various practice conditions. There is a prevailing belief among educators that repeated practice has negative effects on student attitudes and motivation (Bennett, Finn, & Cribb, 1999). That belief is apparent in popular sayings such as "drill and kill" and is maintained and disseminated by influential writers (e.g., Kohn, 1999). It was interesting to find that we could not unearth a single experimental study demonstrating such a detrimental effect of drill and practice. In fact, the few published studies on the topic report either a lack of effect on attitudes (e.g., Davies, 1972) or positive effects on students' motivation (McDermott & Stegemann, 1987; Smith, 1980; Weitzman, 1964) and on their attitudes toward the subject matter (Mevarech, Siber, & Fine, 1991; Vincent, 1977; Wittman, 1996) and the practice activity itself (Heath-Legg, 1992; Lanchantin, 1991). Moreover, Nathan and Baron (1995) report that students often prefer drill-and-practice programs to other types of computer programs, including games and simulations. In spite of the lack of documented support for the "drill and kill" belief, we decided to assess the impact of repeated practice on students' attitudes toward both the course and the computer-controlled practices. The presence of positive effects or the absence of negative ones would reinforce the social acceptability of teaching methods involving a large amount of practice. If any negative reactions were observed, the design would allow some insight on the specific conditions that had generated them.

Method

Participants

A total of 190 college students who were enrolled in an introductory quantitative methods course constituted the initial pool of participants. They were distributed among eight classes and four instructors from two colleges in the greater Montreal area. A vast majority of the students were in their 1st year of college education, which corresponds in Quebec to a 12th year of education. Twenty-two students abandoned the course before the end of the semester and were excluded from the study. Within each class, the remaining students were matched in terms of abilities (see explanation later) and then randomly assigned to three experimental conditions: (a) mastery learning only, (b) accuracy-oriented overlearning, and (c) fluency-oriented overlearning.

Teaching Materials and Practice Instructions

Four distinct—but largely overlapping—item banks, one for each instructor, were developed and grouped into nine different practice units; they varied in size between 497 and 529 items. They were adapted to the lesson sequence and vocabulary of the two textbooks used (Amyotte, 1998; Grenon & Viau, 1999), as well as to the specific topics covered by each instructor. Items on specific topics not covered by an instructor were removed from the item bank of that instructor's students. The items were designed to provide practice for various types of skills, including declarative, procedural, and conditional knowledge (Anderson, 1983). We used task analysis techniques to break down complex problem-solving skills into smaller steps and prerequisite skills; we then designed the questions to be used to practice each of those components separately or in small sequences of a few steps each (Engelmann & Carnine, 1982). The practice activities were controlled by the PracticeMill computerized flashcard software (Péladeau, 2000). The software was installed on the colleges' computer network and was accessible 7 days a week from numerous computer rooms. The computer program collected data on each student, including date, time, and duration of practice, total score obtained on each practice unit, and answers and response time for each item. The practice software also controlled students' access to practice files and provided them with different practice instructions based on their experimental condition.

The instructors were asked to include the computerized weekly practices in their course activities and were invited to give students a minimum of 15 points (out of 100) for attending those practice sessions and attaining the weekly practice criterion. The number of points they chose to give ranged from 5 to 20. They were free to choose whether to schedule the practice periods during or outside class hours. All of the teachers adopted a mixed solution, scheduling some practice time during class hours but asking students to do most of their practice on their own. As soon as an instructor had completed a given topic, the related practice unit became available. Students were initially instructed to practice each available unit at least once weekly until they reached an accuracy criterion of 85% on it or until they made five consecutive unsuccessful attempts.

Experimental Conditions

After the first 2 weeks of computer practice, students in each class were ranked according to their prior ability level and initial performance on the first two practice units. The high school GPA (grade point average) score used by Quebec colleges for student selection became our measure of prior ability. Computed from the scores obtained by students during their last 2 years of high school, it was found to be a very good predictor of college academic performance, with correlations between .65 and .67 (Terrill & Ducharme, 1994). A scholastic aptitude test was administered to 34 students whose GPA scores were not available (Otis, 1962). Scores on that particular academic aptitude test have also been found highly predictive of high school performance, with correlations ranging from .50 to .57 (Gagné & St Père, 2002; Otis, 1962). Within each class, trios of matched-ability students were created and randomly assigned to one of the following experimental conditions:

1. *Mastery Group.* Students in the mastery condition had to practice a specific unit until they reached a stable level of 85%. This level was attained when the last performance in a specific week was above 85% and was followed the next week either by an initial performance at least as good or by an average of at least 85% on the first two trials. When the student reached that criterion for a given practice unit, it was automatically removed from his or her weekly practice load and made unavailable for the remainder of the semester.
2. *Overlearning Group With Accuracy Instructions.* After attaining the same mastery criterion as the first group, members of this first overlearning group were instructed to continue practicing the mastered unit twice a week for up to 5 weeks, focusing on the maintenance or further increase of their accuracy levels. We call this group the "high-accuracy group."
3. *Overlearning Group With Fluency Instructions.* After reaching the initial mastery criterion, students in the second overlearning group had to continue practicing each mastered unit twice a week for 5 consecutive weeks. The number of practice sessions and items was controlled to ensure that the students received the same amount of practice as those in the high-accuracy group. But, unlike the high-accuracy group, these students were asked to focus on increasing their correct response rate. At the beginning of a weekly practice on

mastered units, they were given their highest frequency score, expressed in number of correct responses per minute, and were instructed to increase that response rate. Their performance on each practice unit was expressed in the same response-rate measure. We call this group the "fluency group."

All of the students were informed that a study was being conducted on the effect of computerized practice and that various practice conditions were being tested. They were not told about the specific nature of the variations or the existence of various experimental groups within each class. Students could withdraw from the research at any time; but because computerized practice was an integral part of the course, they were to continue fulfilling the weekly practice requirements set by their instructor. Every week, the students received a list of practice units, each with its performance criterion for that week. They were not told in advance what performance level would bring a given unit to be judged mastered and thus removed from their practice load. The instructors were informed of the three experimental conditions and of the fact that those conditions would produce individual discrepancies in amount of practice. However, they were not told which students belonged to each experimental condition.

Special Comparison Group

Of the 168 students who completed the course, only 54% (n = 90) practiced enough to reach the minimal mastery criterion needed to be included in one of the three experimental groups. At the end of the experiment, we decided to create a special comparison group with these noncomplying students. Major differences in the practice contingencies and classroom management behavior of the four instructors accounted in large part for the high percentage of nonparticipation. Indeed, the percentage of students who practiced enough to reach the initial mastery criterion ranged from 33% for the most permissive instructor to 85% for the instructor who awarded the highest number of points and maintained the most constant pressure to participate. We also found a moderate relationship between the prior ability level of the students and their amount of practice. The ability level of the nonparticipating students was 0.78 standard deviation below that of the students who remained in their experimental group, with 95% *CI* (confidence interval), CI_{95} = (0.46, 1.12). There was also a clear interaction between the students' aptitudes and the instructors' practice contingencies. High-ability students generally participated regardless of the practice contingencies, whereas lower-ability students were much less prone to participate when there was less incentive or pressure to do so. Attrition and failure to practice enough were not related to the various treatment conditions.

Despite a clear self-selection bias that made students in the comparison group nonequivalent, we decided to supplement the study with the quasi-experimental data gathered from this group and to try to control, both statistically and experimentally, for the observed initial differences in ability level. Also, because most students in the comparison group profited from some practice and from the potential benefit of continuous assessment of their performance, they cannot be considered equivalent to a nontreatment condition. Yet the inclusion of such a group allowed us to measure the additional benefit of achieving mastery, as well as to shed some light on the factors that affect student participation.

Dependent Measures

The dependent variables were measured on two separate occasions. The academic performance of all students was assessed during the course and at its end; long-term retention and attitudes toward the course, the subject matter, and computerized practice were measured on a subsample of students about 6 months after the end of the course.

Academic Performance

The impact of various practice conditions on knowledge application was assessed by comparing the performance of the students on their respective within-semester and final exams. The exams, prepared independently by each instructor, consisted mostly of application problems; the students were asked to perform various complex tasks on sample datasets, such as constructing a frequency table, using graphics to illustrate a distribution, computing chi-square values, and performing and interpreting hypothesis tests. When questions from the computerized practice units were used in an exam, those items were excluded from the computation of the exam score.

Retention

To assess whether the three experimental conditions could produce differences in long-term retention of practiced skills, we administered a retention test to a subsample of 83 students. The test consisted of 44 questions covering topics introduced in the first half of the course. We restricted the retention test to those topics because they were the ones on which there were clear differences in practice conditions between the mastery group and the two overlearning conditions. We excluded topics for which most students had not had enough time either to reach the mastery level or to engage in sufficient overlearning trials. Twenty-six of the 44 questions were taken directly from the pool of questions that the students had practiced; 18 new questions were written, similar in format but involving different examples or values. By including these two types of questions we could verify whether any observed differences could be attributed to a simple recall of responses to specific questions or to successful transfer to new examples. This retention test was given 5 to 6 months after the students had ceased practicing the topics covered.

Attitudes

A short, custom-made attitude questionnaire was administered to the 83 students who participated in the retention test. It consisted of seven items wherein students were asked to respond on a 5-point Likert scale ranging from "strongly disagree" to "strongly agree." The questions covered attitudes toward the course and the teaching (three items), statistics as a subject matter (two items), and the computerized practice itself (two items). An open-ended question allowed the students to express in writing their opinion on any topic related to the course, the computerized practice, or the experiment.

Retention Test Procedure

All of the students in the three experimental groups were contacted by telephone and asked to participate in a study on the teaching effectiveness of quantitative methods. No reference was made to the previous study. The students were scheduled to take part at a specific time and location at their college and were offered $20 each for their participation. Of the 89 students we were able to reach by telephone, only 5 refused to participate. Twenty students from the special comparison group were also invited to participate in the retention posttest. The following criteria were used to select students from this group: (a) those with the highest ability level, (b) those who had successfully completed the course, and (c) those who were enrolled in the classes of the two instructors with the lowest participation rates. Selecting only the best students in the non-mastery group considerably reduced the difference in initial ability level between them and those in the three experimental conditions. Also, because the majority of the students who had failed the course were required to take it again the following semester, we had to eliminate currently enrolled students. The decision to restrict the selection of non-mastery students to the four classes with the lowest participation rates was made to ensure that the sample would include students most likely to have participated if they had been provided with stronger incentives.

The experimenter or his research assistant administered the retention test and attitude survey. Students were asked to sign a consent form describing the purpose of the follow-up study. The attitude survey was administered, followed by the paper-and-pencil retention test. The students were told that the purpose of that test was to assess their retention of the material that they had learned during the course. It was emphasized that this test was not intended to evaluate them but to compare the effectiveness of different practice conditions that had been experimented with during the previous semester. They were told that they had 20 minutes to complete the questionnaire but that they should have plenty of time to answer all questions.

Results

This section contains three parts. First, we will discuss technical problems associated with the analysis of the data that we collected. Next, we will look at various indexes of treatment implementation. Finally, we will present the results themselves, organized according to the type of dependent variable and the particular hypothesis tested.

Data Analysis Problems

Planning the data analysis meant facing two major methodological decisions: (a) how to treat the first dependent variable, and (b) which statistical approach to use.

The First Dependent Variable: Measurement of Academic Achievement

To increase ecological validity, we decided to assess the academic impact of the three experimental conditions by using the teachers' own exams rather than a single test specifically designed for the study. Ideally, we hoped to use final exams prepared by each instructor and covering topics from the whole semester. Unfortunately, only one of the instructors used cumulative "final" exams; the others gave exams that covered only the topics taught since the preceding exam. So we decided instead to use the 12 exams (3 per instructor) that were administered over the whole semester at intervals of about 5 weeks. That decision created new problems. First, the fact that we could not assume equal difficulty and content coverage among instructors prevented our creating a semester average for each student and using it for treatment comparisons. Second, even had that been possible, large class differences in participation rates prevented the aggregation of student scores into treatment means because we would then have confounded treatment effects with implementation effects—namely, the impact of instructors on student participation through their differential encouragement and reinforcement of computerized practice. For these reasons, we decided to analyze the differences in exam scores for each instructor separately and then summarize the overall effect with meta-analytic techniques (Schmidt & Hunter, 1990). This strategy had the additional benefit of providing information about the generalizability of the different practice conditions, by allowing a direct assessment of the stability of results across instructors.

Another problem emerged when we examined the exam scores more closely. The practice database revealed that it took typically 2 to 4 weeks for the students to master specific units and an additional 2 to 4 weeks for us to observe clear differences between the mastery-only condition and the two overlearning conditions. Thus the lack of cumulative exams reduced the chances that overlearning effects would reveal themselves, because each exam was administered about 5 weeks after teaching began on a given set of topics. To minimize that problem, we adopted a different strategy for each exam. In the case of the eight control exams within the semester, we included in each treatment group only the students who had achieved mastery for at least half of the units taught during the first half of the course. In the case of the three noncumulative final exams, we included only the students who had mastered at least half of the units introduced in the second half of the course. Finally, the single cumulative exam made it possible to include all of the students in the different experimental conditions.

Statistical Approach: Confidence Intervals Versus Significance Testing

We assessed the impact of the various experimental conditions on academic performance, retention, and attitudes through bivariate regression analyses. When statistical control for initial differences was required, we conducted hierarchical multiple regression analyses or partial correlation analyses. We transformed the resulting correlation coefficients into Cohen's *d* measures of effect size (Cohen, 1988) and reported the effect sizes along with a 95% confidence interval. Several researchers have criticized significance testing; they propose that effect size measures be reported either in addition to or in place of those tests (Cohen, 1994; Morrison & Henkel, 1970; Oakes, 1986; Schmidt, 1996). These repeated criticisms have led to changes in the fifth edition of the *Publication Manual of the American Psychological Association* (American Psychological Association, 2001) concerning data analysis and presentation. The manual notes that, "because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended" (p. 22). In conformity with that trend, with similar recommended changes in the editorial policy of the American Educational Research Association (Thompson, 1996), and with the statistical practices of educational researchers (Keselman, Huberty, Lix, Olejnik, Cribbie, Donahue, et al., 1998), we have chosen to present and interpret effect sizes and confidence intervals.[1] We will use as our interpretative guidelines Cohen's definitions of small, medium, and large effect sizes, which correspond to differences of at least 0.2, 0.5, and 0.8 standard deviation, respectively. For purposes of comparison, Lipsey and Wilson (1993) examined 181 meta-analyses on various educational questions and reported a median effect size of 0.39. Hattie, Briggs, and Purdie's (1996) synthesis of 304 meta-analyses on educational interventions yielded an effect size of 0.40.

Implementation Measures

Table 1 presents descriptive measures for variables designed to assess (a) the initial equivalence of the three experimental samples and the comparison group, and (b) the integrity of the implementation.

Group Equivalence

Mean ages of students in the various experimental conditions were very similar, ranging between 17.4 and 17.6 years. Their sex appears related to participation level. Female students represent 59% of the total sample, but their percentage rises to 68% or 70% in the experimental groups and falls to 45%

Table 1
Comparison of Experimental Groups on Demographic Variables, Prior Academic Achievement, and Treatment Implementation

	Non-mastery		Mastery		High accuracy		Fluency	
Indicator	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*
Initial differences								
Age	17.6	(0.9)	17.5	(1.8)	17.5	(1.4)	17.4	(0.9)
Sex (% of females)	45.2	(0.53)	67.7	(0.48)	67.9	(0.48)	70.0	(0.47)
High school GPA score	78.0	(9.9)	86.5	(10.7)	88.8	(12.5)	87.3	(12.6)
Initial accuracy (%)	67.0	(12.1)	74.9	(7.5)	72.3	(9.8)	70.9	(9.9)
Treatment implementation								
Number of trials	33.0	(44.9)	53.8	(27.2)	100.5	(42.9)	91.5	(31.1)
Total acceleration	2.1	(1.2)	2.3	(0.8)	3.4	(1.0)	4.0	(1.2)
Final accuracy (%)	77.5	(11.1)	89.0	(2.0)	91.1	(4.7)	87.8	(6.1)
Sample size	$n = 78$		$n = 31$		$n = 29$		$n = 30$	

in the non-mastery condition, $d = 0.44$, $CI_{95} = (0.13, 0.77)$. However, no noticeable sex differences appeared between the three experimental groups. The potential impact of these sex differences on comparisons involving the non-mastery condition with any of the three experimental conditions was assessed and found negligible. Sex differences never accounted for more than 0.2% of the total variance of any of our dependent variables. As mentioned previously, the data in Table 1 show substantial differences in prior academic ability between students who did not participate (non-mastery group) and those who did. The differences are found on the measures of both high school GPA, $d = 0.87$, $CI_{95} = (0.54, 1.22)$, and initial performance on each practice unit, $d = 0.54$, $CI_{95} = (0.19, 0.90)$. When relevant, the analyses that follow will statistically control for these important differences. The three experimental groups were similar in terms of prior academic ability. However, the initial performance level of the mastery group students was slightly superior to that of students in the two overlearning conditions, $d = 0.35$, $CI_{95} = (-0.11, 0.84)$.

Treatment Implementation

The number of practice sessions represents the most tangible measure of treatment implementation. Table 1 shows that the two overlearning groups ended up with about twice as much practice as the students in the mastery condition and three times as much as the students who did not reach the mastery criterion. The present study also aimed at establishing differences in response behavior between the two overlearning conditions, with the fluency group showing a faster response rate and the accuracy group showing a higher percentage of accurate responses. The data confirm this hypothesis. The students in the fluency group were able to multiply their initial correct response speed by a factor of 4.0, whereas those in the high-accuracy condition multiplied their initial rate by a factor of 3.4, $d = 0.58$, $CI_{95} = (-0.02, 1.21)$, despite a slightly higher number of trials. The latter group also achieved moderately more accurate performances than the fluency group (91.1% as opposed to 87.8%), $d = 0.61$, $CI_{95} = (0.02, 1.24)$.

The large group differences on the final accuracy measure with respect to within-group variability (*SD*) may also reflect differences in implementation conditions. The data in Table 1 show a standard deviation more than twice as large for the high-accuracy group as for the mastery group, and more than three times as large as for the fluency group. These differences are consistent with the fact that the accuracy level was used as the termination criterion for the mastery group, creating a floor effect in the distribution of that variable. On the other hand, the two overlearning groups kept practicing for weeks, with performances that sometimes fell well below the 85% mastery criterion, thus accounting for greater variability. The slightly larger variance that was observed in the fluency condition, when compared with the other overlearning group, is also consistent with the experimental instructions. Our examination of the performance profiles of individual students in the fluency group revealed that the instruction to focus on response rate was often followed by a decrease in accuracy; it appeared that the students were trying hard to respond more quickly, often at the cost of accuracy.

Treatment Effects

This section presents the results on the impact of the various practice conditions; it is structured according to the three dependent variables: short-term academic performance (labeled *academic achievement*), long-term retention (labeled *retention*), and attitudes toward the learning environment (labeled *attitude*).

Academic Achievement

One of the main goals of any teaching activity is to help students to better understand the subject matter and ultimately to succeed. The first research question addressed the positive impact of building mastery. By far, the most impressive difference observed between the non-mastery condition and the other three groups was directly related to group failure rate. The introductory course in quantitative methods had one of the highest failure rates of all the compulsory core courses in the social sciences college program (Commission d'évaluation de l'enseignement collégial, 1995). Usually, more than one third of the students who completed the course failed to reach the minimum passing grade. Our results were comparatively much more satisfactory, with a global failure rate of only 20% (33/168). Moreover, with only one exception, all of those who failed belonged to the non-mastery group, raising that group's failure rate to 41% (32/78). This huge difference still holds when we take into account the prior academic abilities of the students, $d = 0.87$, $CI_{95} = (0.55, 1.22)$. Based on the relationship between GPA and achievement *within* the non-mastery group, we estimated that the failure rate in the three experimental groups should have been 27%. The fact that it was close to zero speaks eloquently for the value of mastery learning. This finding is consistent not only between groups but also between classes. There was a direct relationship between the participation rate induced by the various instructors and the failure rates observed in their classes. The instructor whose students participated least (33%) had the highest failure rate (35% of those who completed the course); the one with the highest level of participation (85%) obtained the lowest failure rate (2%). A regression analysis between the failure rate observed in the eight classes and their participation rate confirms this strong relationship, $r = -.91$, $CI_{95} = (-.98, -.55)$.

These results leave room for an alternative explanation. Recall that up to 20 points could be obtained simply by practicing. Now, the most successful instructors—those with the lowest failure rates—were also the ones who had given the highest number of points for practice. It makes sense that the absolute number of points attributed could have made the difference between success and failure for more students of the "generous" instructors, thus explaining the observed class discrepancies in failure rate. To rule out this alternative explanation, we need to demonstrate that the computer-controlled practice also improved the students' performance on other course activities, especially on the various exams administered by the instructors. As mentioned earlier, effect sizes on exam scores were computed separately for all 12 exams and then aggregated into an overall effect size using meta-analytic statistical procedures (Schmidt & Hunter, 1990). Table 2 presents these aggregated effect size values. Moreover, for all comparisons involving the non-experimental group, effect sizes were computed both on the original scores and on scores statistically adjusted for differences in prior academic ability. The unadjusted differences between the non-mastery group and the three experimental groups confirmed that practicing up to mastery or beyond increased exam scores by a margin of 0.56 to 0.83 standard deviation above the performance of students who did not reach the mastery criterion. The adjusted effect sizes suggest that part of this difference may be attributed to the observed initial ability discrepancies; however, practice still produces a modest positive impact, with aggregated effect sizes ranging between 0.16 and 0.43. To interpret these results correctly, readers must understand the difficulty in assessing the respective contributions of (a) the experimental conditions, and (b) student academic aptitudes; the difficulty ensues mainly from the moderate correlation between level of participation and these abilities. The percentage of explained variance attributed to the treatment drops from 22.8% to 7.3% when we control for prior academic abilities. Alone, that variable explains 31.7% of the observed variance on exam scores. However, because students with poor academic aptitude were also less likely to partic-

Table 2
Differences Between Experimental Conditions on Pooled Periodic Exams and Final Exams, in *SD* Units (95% Confidence Intervals)

Academic performance	Non-mastery	Mastery	High accuracy
Mastery			
Unadjusted effect size	0.56 (0.34, 0.78) p = .000		
Adjusted effect size	0.31 (0.09, 0.53) p = .006		
High accuracy			
Unadjusted effect size	0.83 (0.50, 1.09) p = .000	0.57 (0.25, 0.91) p = .000	
Adjusted effect size	0.43 (0.19, 0.67) p = .000		
Fluency			
Unadjusted effect size	0.57 (0.33, 0.83) p = .000	0.42 (0.05, 0.76) p = .025	−0.03 (−0.49, 0.39) p = .878
Adjusted effect size	0.16 (−0.08, 0.40) p = .201		

Note. Effect sizes in this table result from a meta-analytic synthesis of effect sizes computed on twelve control exams. Confidence intervals are given in parentheses.

ipate, it is reasonable to assume that an unspecified amount of the common variance assigned to this controlling variable could be attributed to differences in level of practice. In other words, although effect sizes computed without controlling for prior academic abilities clearly overestimate the real benefit of the practice conditions, their statistical control yields effects sizes that underestimate the impact of practice. The two values thus represent upper and lower limits between which the real effect sizes probably lie.

The separate analysis of the exam scores for each class of each instructor offers the additional advantage of establishing the consistency of the observed differences and, consequently, the generalizability of the conclusions based on aggregate measures. The academic benefit of reaching mastery performance (with or without overlearning) emerged in a large majority of the exams. In fact, only 3 of the 36 effect sizes (e.g., 4 instructors × 3 exams × 3 comparisons) that were computed with unadjusted scores between the non-mastery condition and the three other experimental conditions were in the opposite direction. The ratio increased to 6/36 when we used adjusted scores.

The second research question was whether overlearning provides any additional benefit. The fact that only one student failed in all three experimental groups keeps us from using that performance indicator. However, comparisons of exam scores between those who merely reached the mastery criterion and those who engaged in overlearning show that the latter performed slightly-to-moderately better, with their scores exceeding those of the mastery group by 0.42 to 0.57 standard deviation (see Table 2). This effect was highly consistent across instructors and exams. Students in both overlearning conditions outperformed mastery group students on 21 of the 24 comparisons. Concerning the third research question, namely, expected differences between the two overlearning conditions, our results indicate that fluency instructions did not increase academic achievement beyond standard overlearning practices: Both groups obtained almost identical scores, $d = -0.03$. This finding is also reflected in an instability of observed effect sizes computed on separate exams; we observed an equal proportion of negative and positive effect sizes, confirming the absence of any systematic difference between these two conditions.

Two final observations need to be made about the exam results. If we exclude comparisons between the two overlearning conditions, 7 of the 9 negative results observed among the remaining 60 comparisons came from a single instructor. It is also worth noting that the only cumulative final exam produced the most important beneficial effects. Not only did the mastery group surpass the non-mastery group [adjusted standardized difference = 0.51, CI_{95} = (–0.40, 1.52)], but also the performance of the two overlearning groups exceeded that of the mastery group by no less than an additional 1.15 standard deviation, CI_{95} = (–0.04, 2.59). However, these two effect sizes come from very small samples ($n = 23$ and $n = 18$); because of their limited reliability, the estimations should be considered tentative.

Retention

For the retention posttest, the selection of the best students from the non-mastery group (see "Method" section) reduced the large comparative weakness of the latter group in terms of prior academic ability and initial performance on computer-controlled practice. In comparison with the students in the three experimental conditions who also participated in this retention test, this selected group had only slightly lower GPA scores (85.9 as opposed to 89.1), $d = 0.24$, CI_{95} = (–0.20, 0.69), and almost identical initial comprehension scores (70.6% as opposed to 72.9%), $d = 0.11$, CI_{95} = (–0.33, 0.55). Because statistical control for the remaining differences had no noticeable impact on the main observations, we used unadjusted effect sizes. Table 3 presents descriptive statistics for the total score on the retention test, as well as for recall and transfer items separately. Bivariate comparisons between various subgroups on these measures appear in Table 4. With respect to the first hypothesis on the effect of attaining mastery, the performance gap shown in Table 3 between the non-mastery and mastery groups illustrates the important benefit of reaching mastery. The mean total retention score of students in the first group was 55.7% as opposed to 65.3% in the mastery group, $d = 0.77$, CI_{95} = (0.13, 1.46).

Concerning the second hypothesis, about the added benefits of overlearning, students in both overlearning conditions performed much better than those in the mastery group. The two overlearning groups obtained

Table 3
Scores Obtained on Retention Test and Attitude Test for All Experimental Conditions

Follow-up measure	Non-mastery		Mastery		High accuracy		Fluency	
	M	*(SD)*	*M*	*(SD)*	*M*	*(SD)*	*M*	*(SD)*
Retention								
Total score	55.7%	(11.3)	65.3%	(13.9)	75.2%	(9.1)	75.0%	(9.8)
Recall	55.6%	(11.2)	66.1%	(16.1)	73.3%	(10.4)	75.3%	(12.3)
Transfer	55.8%	(15.8)	64.1%	(14.1)	78.7%	(9.8)	74.7%	(9.5)
Attitude								
Global	3.18	(0.49)	3.44	(0.51)	3.55	(0.51)	3.55	(0.61)
Course	2.90	(0.70)	3.17	(0.74)	3.19	(0.76)	3.24	(0.82)
Statistics	3.40	(0.94)	3.57	(0.79)	3.82	(1.15)	3.48	(1.05)
Practice	3.37	(0.83)	3.72	(0.86)	3.82	(0.61)	4.10	(0.62)
Sample size	$n = 20$		$n = 23$		$n = 19$		$n = 21$	

almost identical retention scores of 75.2% and 75.0%, yielding effect sizes of 0.84 and 0.82, respectively. The cumulative benefits of reaching mastery and engaging in overlearning trials (see Table 4) is expressed by very large differences between the non-mastery group and both the high-accuracy group ($d = 1.95$) and the fluency group ($d = 1.88$). The retention data did not support the third hypothesis: Whether students who engaged in overlearning were instructed to maintain and further increase their accuracy or to focus on improving their response rate did not influence long-term retention. Indeed, as shown above, both groups obtained similar high retention scores ($d = -0.02$). Another substantial difference is worth noting: Not only did the students in the two overlearning conditions achieve much better scores, but they also did so in less time than the other two groups. They answered all 44 questions in an average of 10.7 minutes, as compared with 12.8 minutes for the other two groups, $d = -0.74$, $CI_{95} = (-1.22, -0.29)$, which suggests that students in both overlearning conditions attained higher proficiency.

To assess the specific contribution of prior academic achievement and of the various practice conditions on retention, we performed a covariance analysis. Prior academic abilities, which we entered first in the statistical model, accounted for 18.3% of the total variance of retention scores. The experimental conditions explained an additional 28.4% for a total of 46.7% for these two variables alone. It is important to stress that by using prior academic abilities to select students for the non-mastery condition we somewhat reduced the variability of this measure and, consequently, its ability to predict retention. However, that selection reduced the common variance between the two predictors and made it possible to better estimate the specific contribution of the experimental conditions on retention.

Table 4

Differences Between Experimental Conditions on Retention Test and on Attitude Test, in *SD* Units (95% Confidence Intervals)

Follow-up measure	Non-mastery	Mastery	High accuracy
Retention test			
Mastery			
Total score	0.77 (0.13, 1.46) $p = .019$		
Recall	0.76 (0.12, 1.45) $p = .019$		
Transfer	0.57 (−0.06, 1.24) $p = .076$		
High accuracy			
Total score	1.95 (1.11, 2.94) $p = .000$	0.84 (0.19, 1.56) $p = .011$	
Recall	1.68 (0.89, 2.60) $p = .000$	0.53 (−0.10, 1.20) $p = .100$	
Transfer	1.63 (0.89, 2.59) $p = .000$	1.11 (0.43, 1.88) $p = .001$	
Fluency			
Total score	1.88 (1.07, 2.82) $p = .000$	0.82 (0.19, 1.52) $p = .011$	−0.02 (−0.67, 0.63) $p = .951$
Recall	1.72 (0.94, 2.62) $p = .000$	0.65 (0.03, 1.32) $p = .040$	0.18 (−0.46, 0.84) $p = .584$
Transfer	1.49 (0.75, 2.35) $p = .000$	0.89 (0.25, 1.60) $p = .006$	−0.34 (−1.00, 0.31) $p = .308$
Attitude test[a]			
Mastery			
Global attitude	0.71 (0.07, 1.39) $p = .028$		
Course	0.75 (0.11, 1.44) $p = .024$		
Statistics	0.43 (−0.19, 1.08) $p = .187$		
Practice	0.09 (−0.53, 0.71) $p = .786$		
High accuracy			
Global attitude	0.96 (0.27, 1.72) $p = .006$	0.20 (−0.43, 0.84) $p = .548$	
Course	0.63 (−0.04, 1.34) $p = .014$	−0.01 (−0.65, 0.62) $p = .964$	
Statistics	0.40 (−0.26, 1.08) $p = .246$	0.26 (−0.36, 0.91) $p = .422$	
Practice	0.44 (−0.21, 1.13) $p = .197$	0.13 (−0.50, 0.77) $p = .685$	

(continued)

Table 4 (Continued)
Differences Between Experimental Conditions on Retention Test and on Attitude Test, in *SD* Units (95% Confidence Intervals)

Follow-up measure	Non-mastery		Mastery		High accuracy	
Fluency						
Global attitude	0.66	(0.01, 1.36) $p = .051$	0.20	(−0.42, 0.82) $p = .539$	0.02	(−0.63, 0.67) $p = .959$
Course	0.71	(0.06, 1.42) $p = .036$	0.09	(−0.53, 0.70) $p = .787$	0.11	(−0.54, 0.76) $p = .749$
Statistics	0.12	(−0.52, 0.76) $p = .720$	−0.10	(−0.71, 0.52) $p = .765$	−0.32	(−0.99, 0.32) $p = .337$
Practice	0.52	(−0.13, 1.19) $p = .126$	0.53	(−0.09, 1.18) $p = .104$	0.47	(−0.18, 1.15) $p = .170$

Note. Confidence intervals are given in parentheses.
[a]Comparisons on the attitude test were performed on scores adjusted for observed differences between instructors.

Attitude

Descriptive statistics for the overall attitude score and the three subscales (attitudes toward the course, statistics, and computerized practices) are presented in Table 3. It is noteworthy that the most positive means belong to items related to computerized practice. The combined attitude score on this subscale was 3.8, as compared with an average score of 3.6 for attitudes toward statistics and 3.1 for attitudes toward the course. Eighty-six percent of the students felt that computerized practice helped them to better understand the subject matter. Comparisons among students from different classes yielded important differences, especially with regard to attitudes toward the course and toward practice. Students in the classes of instructors who had induced higher participation rates showed more positive attitudes toward this activity, $r = .32$, $CI_{95} = (.12, .50)$. The bottom half of Table 4 shows comparison results on those measures statistically adjusted for the observed differences between instructors in terms of student satisfaction expressed. The data reveal that participating students showed more positive attitudes than nonparticipating students. These differences in attitude are especially large and consistent in the case of the global score, $d = 0.66$, and the course subscale, $d = 0.96$. The non-mastery and mastery groups did not differ substantially in their attitudes toward computer-controlled practice, but students in both overlearning conditions showed slightly to moderately more positive attitudes toward the practice activity. It is interesting that students who were instructed to focus on increasing their response rate expressed more positive attitudes toward practice than did both the students who stopped at initial mastery, $d = 0.53$, $CI_{95} = (-0.09, 1.18)$, and those in the high-accuracy overlearning condition, $d = 0.47$, $CI_{95} = (-0.18, 1.15)$. In short, contrary to what one would expect on the basis of popular beliefs, we were unable to document any detrimental motivational effect of engaging students in overlearning trials beyond initial mastery. All differences between the high-accuracy group and the mastery group were negligible, ranging from −0.01 to 0.26 standard deviation, most of them in favor of the overlearning condition. The only potential indication of a negative impact was found for the fluency group students, who showed slightly more negative attitudes toward statistics than did the students in the high-accuracy overlearning condition, $d = -0.34$, $CI_{95} = (-0.99, 0.31)$. However, their attitudes toward the subject matter did not differ from those of students in the mastery and non-mastery conditions.

Discussion

The main purpose of this study was to examine the impact of various amounts of practice and various practice conditions on academic performance and long-term retention. More specifically, we tried to verify whether fluency-building practices would provide additional benefits in comparison with the standard overlearning exercises in which students typically are asked to maintain their accuracy level.

Summary of Findings

Comparisons between students who did and did not reach the mastery criterion lead us to conclude that practicing until mastery does provide a substantial benefit, whether we look at exam scores, course failure rate, or long-term retention. In this respect, our results support previous findings, which show that practicing until mastery has a positive impact on academic achievement (Block & Burns, 1976; Guskey & Gates, 1986; Kulik, Kulik, & Bangert-Drowns, 1990; Kulik, Kulik, & Cohen, 1979; Slavin, 1987) and long-term retention (Farr, 1987). Comparisons between students in the two overlearning groups and those in the mastery condition reveal that engaging students in overlearning trials provides further benefits, especially with respect to long-term retention. These positive results from an applied setting extend earlier findings from laboratory studies on memory (Driskell et al., 1992; Farr), bringing much-needed ecological validity to this area of research. Moreover, contrary to Driskell et al.'s observation that the effect of overlearning lasts only 5 or 6 weeks, we observed large skill differences 5 to 6 months after the students had stopped practicing.

With respect to another important research question, namely, the differential effectiveness of standard overlearning trials in comparison with paced practice focusing on an increase in response rate, our results show that students in the paced practice condition were able to attain slightly higher response rates than those in the standard overlearning condition; but that difference affected neither academic achievement nor long-term retention. This lack of effect replicates some prior research (Makepeace, 1998; Ormrod & Spivey, 1990; Pedego, 1999) but contradicts the conclusions of other experimental studies (Berquam, 1981; Kelly, 1996). Finally, the results of the current study provide no support for the alleged demotivating effect of drill and practice. Our data indicate instead that repeated practice may help to improve attitudes toward (a) the course, (b) the subject matter, and (c) computerized practice. The positive reactions of our participants confirm the conclusions of many previous studies, namely, that students often prefer drill-and-practice software over other types of educational software (Nathan & Baron, 1995) and that this learning technique often produces positive affective outcomes (Heath-Legg, 1992; Lanchantin, 1991; McDermott & Stegemann, 1987; Mevarech et al., 1991; Smith, 1980; Vincent, 1977; Weitzman, 1964; Wittman, 1996). In fact, during the retention-building phase, focusing the practice on an increase in correct response rate rather than on the maintenance of high accuracy levels may increase students' interest in the task and, consequently, their general satisfaction with the course. Pedego similarly observed that students appear to prefer faster fluency-based practices to slower overlearning exercises. This finding also provides empirical support for Lindsley's (1996) observation that fluency training is enjoyable.

Still, the specific nature of the course could explain part of that positive attitude toward computerized practice. As mentioned earlier, the quantitative methods course has the well-earned reputation of being one the most difficult courses in the social sciences program. Consequently, one would expect students to experience more anxiety and fear of failure in relation to this course. Students might have perceived as a welcome relief the academic support provided by repeated practice and mastery learning techniques.

Validity Issues

To interpret these findings appropriately, we need to take into account both the strengths and the weaknesses of our methodology. From the beginning, we rejected the option of a controlled experiment in a laboratory setting. Instead, we asked college instructors to incorporate a weekly practice activity into their existing course. The great deal of effort spent in assessing the implementation process and the effects of these teaching techniques in a regular college setting helped to reinforce the study's ecological validity; it also permitted us to collect valuable information about implementation problems. Moreover, the efforts invested in the methodological design made it possible to circumvent two of the major weaknesses of applied studies in education, namely, instructor and implementation biases. We avoided instructor bias by performing pairing and random assignments in each class and then letting the software control the administration of experimental conditions to different students in the same class, thus approximating a double-blind control situation. Because the instructors did not know to which experimental condition each student belonged, their potential attitude differences toward the various conditions could no longer bias their grading behavior.

Implementation bias ensues from the usual lack of any data on the quality of the implementation process. Applied studies often fail to produce statistically or practically significant results simply because many of the individuals assigned to the experimental groups are never properly exposed to the experimental conditions (Cook & Poole, 1982; Rossi, Freeman, & Lipsey, 1999). By collecting implementation data for each student, we were able to include in the analyses only those who were actually exposed to the various experimental conditions. Similarly, we could select for the retention part of the study only the topics for which we were able to establish clear differences in level of practice between the three experimental conditions. Considering the poor implementation conditions observed in a few classes, we would probably have obtained inconclusive results or trivial differences if we had compared all the students initially assigned to the various experimental conditions.

The creation of a comparison group made of students who did not reach the minimal mastery criterion inadvertently introduced two potential selection biases: differences in academic aptitudes (non-mastery students had lower academic aptitudes) and differences in the contingencies imposed by the various instructors (non-mastery students more often came from classes with fewer points given for practice). We were able to control statistically for these group differences when we analyzed the course failure rates and performance on the within-semester and final exams. It remains possible that these students differed on other unmeasured characteristics, such as initial motivation, aspirations, or interest in the subject matter; such differences would artificially inflate that group's performance gap in comparison with the three experimental groups. For that reason, we believe that the observed effect sizes might overestimate to some extent the real impact of both mastery learning and overlearning in the context of computer-controlled practice. However, it seems unlikely that these selection biases could affect comparisons *between* the three experimental conditions, because failure to practice enough (a) was found to be unrelated to the treatment conditions, and (b) did not result in any noticeable initial differences among the three groups. We also consider the findings on the retention test to be less prone to such selection biases. We took care to choose for the non-mastery condition students who had the highest ability levels and belonged to classes with the lowest participation rates. These selection criteria ensured that the group would be representative of students more likely to have participated if offered better incentives to do so. The criteria also greatly reduced the likelihood that unmeasured initial differences between participating and nonparticipating students could explain the observed differences.

While the selection of students and topics for the retention posttest contributed to the study's internal validity, it could also affect the external validity of the findings, especially with regard to the observed effect sizes on the retention test. If college instructors were to apply a similar teaching technique in their own classes, few would observe performance improvements or retention increases as large as ours, and that for at least two reasons. First, our implementation data suggest that unless instructors reward the computerized practice "generously," only a minority of students will practice enough to reach the initial mastery level. The need for external incentives seems especially important if we consider that the students who are least likely to participate are the lower achievers, precisely those who would benefit most from such practice; our results indicate that high-achieving students adopt computerized practice more readily even under low-incentive conditions. As a consequence, the more a teacher is lenient or relies solely on intrinsic motivation to engage students in computerized practice, the wider the gap will be between low and high achievers. It also seems plausible, as shown by our own data, that for many topics, especially those introduced toward the end of the course, there would not be enough practice opportunities for the students to engage in overlearning or even to reach mastery. This situation represents a practical challenge to educators and instructional designers. Given the benefits of practicing up to and beyond initial mastery, how can teaching materials and practice activities be designed so that even the topics introduced at the end of a course receive the required amount of practice to ensure that they will be mastered and retained? Also, because overlearning practice seems to be especially beneficial for long-term retention, the shorter the delay between practice on a specific topic and the exam on that topic, the more difficult it would be to observe its positive effects on academic performance. Our data tend to support that hypothesis to some extent. Only one of the four instructors administered a cumulative exam covering the content of the entire semester. The comparisons for that specific exam yielded the largest achievement differences between the non-mastery group and both the mastery group and the overlearning group.

The fact that female students participated more and were thus overrepresented in our three experimental groups constitutes another potential threat to the generalizability of our study. The greater amount of time spent by female students on computerized practice is, by itself, not surprising; many studies have shown that female college students spend more time per week studying and doing homework than do their male counterparts (Terrill & Ducharme, 1994). Although this sex difference in participation had no noticeable impact on our results and conclusions, it does raise additional practical issues for instructors, especially with regard to the motivation of male students. Despite their reputation for harboring more positive attitudes toward computers (Kay, 1992), in this study they showed less proneness to engage in repeated computerized practice.

Most of the contradictory findings observed on the exam scores came from the two classes of one instructor; that observation also raises some concerns regarding the generalizability of our findings to any instructor. The low participation rate in his classes and random variations in the assignment process may account for the results; yet we cannot rule out the possibility of some mismatch between the content of the computerized practice and the instruction provided by that instructor, or of an interaction between that instructor's teaching style and the practice requirements.

The Role of Fluency

Our inability to observe any additional benefit for paced overlearning practice over unpaced practice does not negate the potential usefulness of attaining response fluency or automaticity. Although students in the paced condition were able, despite an equal amount of practice, to achieve higher response rates than those in the unpaced overlearning condition, it is not clear whether their faster performance truly reflected a decrease in response latency or whether it was caused by faster reading, word skipping, or more efficient cue selection. One way to partially test these alternative explanations would be to give students in both the paced and unpaced overlearning conditions the opportunity to respond as fast as they can, and then verify whether the observed difference still holds. Our study provides evidence only that the specific pacing instruction that we used produced no academic benefit other than more positive attitudes. Thus the initial question remains unanswered: Assuming that response fluency is a good measure of association strength (Judd & Glaser, 1969; Peterson, 1965), is there an experimental manipulation that would, for an equal amount of practice, reduce this response latency? The fact that some researchers have been able to obtain higher response rates with an equal or even lesser amount of practice (Berquam, 1981; Ormrod & Spivey, 1990) suggests that such manipulations may exist. However, even if such techniques were discovered, their practical value would remain unproven until (a) we can identify specific practice instructions that reliably produce such an effect in an educational setting, and (b) it is demonstrated that such a decrease in response latency produces enhanced retention or transfer.

Even if researchers failed to identify such conditions, behavior rates and response latency could still be useful as measuring devices, allowing one to determine whether specific students need further practice or should move on to the next learning unit. This search for an appropriate criterion of the right amount of practice dates back to the early history of psychological and educational research (Ebbinghaus, 1885/1964; Thorndike, 1921). More recently, Driskell et al. (1992) reaffirmed the importance of further research to identify the point above which the costs of additional practice outweigh the obtained educational benefits. In 1969, Judd and Glaser suggested that "measuring response latency may provide a means of determining the optimal amount of overlearning practice for a particular student and particular group of words" (p. 29). To our knowledge, with the exception of Taymans's (1985) study, no one has pursued this potentially useful avenue. In fact, we must turn to the Precision Teaching literature to find some tentative answers to this question. The huge amount of frequency data obtained through years of teaching efforts by precision teachers has led to the development of normative fluency standards for several tool skills (Haughton, 1980), as well as procedures for the development of performance standards for more complex skills (White, 1985). The question whether fluency standards provide adequate criteria for identifying the right amount of practice, or whether other types of criteria—based on either response latency measures or relative response rates—would be more adequate or useful, invites further empirical research.

We must emphasize that even without any empirical evidence for an effect of paced practice on application or long-term retention, there are at least two reasons why one may want to engage students in paced practice.

First, as we have been able to confirm in this study, focusing practice on the increase of response rate may have a motivational value and help to maintain interest, or at least prevent some loss of interest. Second, as mentioned by Bruce (1999), fluency-building techniques may provide a way to improve learning efficiency. From this point of view, any instructional method that produces an equal amount of practice in less time should be preferred to a less efficient method. This statement parallels an idea expressed by Heward (1994): that the amount of time that students spend on a task is less important than the number of academic responses they make during that time. In this respect, paced practice may allow teachers and students to save precious instructional time, time that could be used for other kinds of activities.

The Instructional Role of Repeated Practice

The bad reputation of drill and practice among educators may be justified to some extent by the prevalence of drill activities that are poorly designed and not related closely enough to important instructional goals. Thus it is important to stress that for repeated practice to be beneficial, the practice items should be carefully designed so that they enable students to practice and learn knowledge and skills that are really prerequisite to the desired terminal performance. Drill does not need to be restricted to the learning of factual knowledge but may also promote the learning of procedural and conditional knowledge and skills. Our own research suggests that class activities should offer enough practice opportunities to ensure the acquisition and retention of specific "tool" skills and knowledge. The practice activities should also provide sufficient contextual variability to facilitate transfer of learning (Cormier & Hagman, 1987). Many educators consider "higher-order" skills too complex to be taught through computerized practice. However, a detailed content analysis of many of these complex skills and a careful examination of common errors made by students when applying them may allow teachers and instructional designers to identify subskills or specific knowledge that could be systematically taught to facilitate the acquisition and mastery of those complex skills. An analysis of situations where transfer did not occur could also reveal the contextual elements (e.g., types of examples and distractors, presentation formats) that need to be included in such instruction. There has been a tendency in education to define educational goals as the acquisition of general thinking abilities that could be developed without reference to any particular content (Bonnett, 1995). The proponents of such goals often consider that drill, repeated practice, and the acquisition of domain-specific knowledge inhibit the development of generic skills such as critical thinking, problem solving, and creativity. However, more and more educators and researchers now recognize that a wide and deep domain knowledge base is an essential condition for the development of expert performance (Glaser & Chi, 1988). Thus the specific contribution of computerized practice to the development of higher-order skills should not be neglected. At the same time, exercises on partial tasks alone may well be as inappropriate as using complex practice settings exclusively. Efforts to master and overlearn basic or tool skills and basic knowledge need to be combined with opportunities for students to engage in more complex activities, both because of their motivational value and because of the specific skills involved in the more complex performances (Anderson, Reder, & Simon, 1996). Focusing exclusively on complex learning situations at the expense of carefully sequenced instruction and repeated practice, as some educators have suggested, may prove to be counterproductive. Our own data suggest that when students are asked to work on complex cognitive tasks, too many of them perform poorly because, for lack of enough practice, they have not mastered, let alone retained, the basic knowledge and skills necessary to complete such tasks. In our view, the promoters of new teaching methods often stress the differences between their own and existing ones; they sometimes even call for the dismissal of valuable teaching techniques, described as incompatible rather than complementary. Over the past 2 decades, methods based on teacher-directed instruction and repeated practice have suffered more than many from such confrontations. The results of our study lead us to believe that most students would benefit from a more balanced set of carefully designed instructions that include repeated practice as well as complex learning activities.

Notes

Correspondence concerning this article should be addressed to Normand Péladeau (see biographical sketches on second page of the article).

[1]We align ourselves with statisticians who hold that testing for significance before looking at effect sizes is incompatible with the logic behind the use of confidence intervals (e.g., Thompson, 2002). They judge the presentation of p values redundant. We include p values in the tables, but we do not take them into account when presenting and discussing the results.

References

Adams, G. L., & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR.* Seattle, WA: Educational Achievement Systems.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Amyotte, L. (1998). *Méthodes quantitatives: Applications à la recherche en sciences humaines* [Quantitative methods: Applications in social sciences research]. Montreal, Quebec, Canada: Édition du Renouveau Pédagogique.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated versus cognitive perspectives: Form versus substance. *Educational Researcher, 26*(1), 18–21.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108,* 296–308.

Bennett, W. J., Finn, C. E., & Cribb, J. T. E. (1999). *The educated child: A parent's guide from preschool through eighth grade.* New York: Free Press.

Berquam, E. M. (1981). The relation between frequency of response and retention on a paired associates task. *Dissertation Abstracts International, 42*(06), 2460A. (UMI No. 8127415)

Binder, C. (1993). Behavioral fluency: A new paradigm. *Educational Technology, 23*(10), 8–14.

Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst, 19,* 163–197.

Binder, C., Haughton, E., & Van Eyk, D. (1990). Increasing endurance by building fluency: Precision Teaching attention span. *Teaching Exceptional Children, 22,* 24–27.

Block, J. H., & Anderson, L. W. (1975). *Mastery learning in classroom instruction.* New York: Macmillan.

Block, J. H., & Burns, R. B. (1976). Mastery learning. *Review of Research in Education, 4,* 3–49.

Bloom, B. S. (1976). *Human characteristics and school learning.* New York: McGraw-Hill.

Bonnett, M. (1995). Teaching thinking, and the sanctity of content. *Journal of Philosophy of Education, 29,* 295–309.

Brophy, J., & Evertson, C. (1976). *Learning from teaching: A developmental perspective.* Boston: Allyn and Bacon.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research in teaching* (3rd ed., pp. 328–375). New York: Macmillan.

Brown, S. A., Dunne, J. D., & Cooper, J. O. (1996). Immediate retelling's effect on student retention. *Education and Treatment of Children, 19,* 387–407.

Bruce, G. S. (1999). *Instructional design made easy: A workbook for designing accelerated learning programs* (2nd ed.). Tucker, GA: Aubrey Daniels & Associates.

Bucklin, B. R., Dickinson, A. M., & Brethower, D. M. (2000). A comparison of the effects of fluency training and accuracy training on application and retention. *Performance Improvement Quarterly, 13,* 140–163.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19,* 237–248.

Commission d'évaluation de l'enseignement collégial. (1995). *Guide spécifique pour l'évaluation des programmes d'études: Le programme de sciences humaines* [Specific guide for the evaluation of programs of studies: The social sciences program]. Quebec, QC, Canada: Author.

Cook, T. J., & Poole, W. K. (1982). Treatment implementation and statistical power: A research note. *Evaluation Review, 6,* 425–431.

Cormier, S. M., & Hagman, J. D. (1987). *Transfer of learning: Contemporary research and applications.* San Diego, CA: Academic Press.

Davies, T. P. (1972). An evaluation of computer assisted instruction using a drill and practice program. *Dissertation Abstracts International, 32*(12), 6970B. (UMI No. 7218627)

Delquadri, J., Greenwood, C. R., Stretton, K., & Hall, R. V. (1983). The peer tutoring game: A classroom procedure for increasing opportunity to respond and spelling performance. *Education and Treatment of Children, 6,* 225–239.

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results. *American Psychologist, 43,* 627–634.

Dougherty, K. M., & Johnson, J. M. (1996). Overlearning, fluency, and automaticity. *The Behavior Analyst, 19,* 289–292.

Driskell, J. E., Willis, R. P., & Cooper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77,* 615–622.

Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology.* New York: Dover. (Original work published 1885)

Engelmann, S., & Carnine, D. (1982). *Theory of instruction: Principles and applications.* New York: Irvington.
Evans, S. S., & Evans, W. H. (1985). Frequencies that ensure skill competency. *Journal of Precision Teaching, 6,* 25–30.
Farr, M. J. (1987). *The long-term retention of knowledge and skills: A cognitive and instructional perspective.* New York: Springer.
Fischer, C. W., Berliner, D. C., Filby, N. N., Marlianve, R., Cahen, L. S., & Dishaw, M. M. (1980). Teaching behaviors, academic learning time, and student achievement: An overview. In C. Denham & A. Lieberman (Eds.), *Time to learn* (pp. 7–32). Washington, DC: National Institute of Education.
Gagné, F., & St Père, F. (2002). When IQ is controlled, does motivation still predict achievement? *Intelligence, 30,* 71–100.
Gillingham, M. G., & Guthrie, J. T. (1987). Relationships between CBI and research on teaching. *Contemporary Educational Psychology, 12,* 189–199.
Glaser, R., & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The Nature of Expertise* (pp. xi–xxxvi). Hillsdale, NJ: Erlbaum.
Greenwood, C. R., Delquadri, J. C., & Hall, R. V. (1984). Opportunity to respond and student academic performance. In W. L. Heward, T. E. Heron, D. S. Hill, & J. Trap-Porter (Eds.), *Focus on behavior analysis in education* (pp. 58–88). Columbus, OH: Charles E. Merrill.
Greenwood, C. R., Hart, B., Walker, D., & Risley, T. (1994). The opportunity to respond and academic performance revisited: A behavioral theory of developmental retardation and its prevention. In R. Gardner III, D. M. Sainaito, J. O. Cooper, T. E. Heron, W. L. Heward, J. W. Eshleman, & T. A. Grossi (Eds.), *Behavior analysis in education* (pp. 213–223). Pacific Grove, CA: Brooks/Cole.
Grenon, G., & Viau, S. (1999). *Méthodes quantitatives en sciences humaines* [Quantitative methods in human sciences] (2nd ed.). Montréal: Gaétan Morin.
Guskey, T. R., & Gates, S. L. (1986). Synthesis of research on the effects of mastery learning in elementary and secondary classrooms. *Educational Leadership, 43*(8), 73–80.
Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research, 66,* 99–136.
Haughton, E. (1972). Aims, growing, and sharing. In J. Jordan & L. Robbins (Eds.), *Let's try doing something else kind of thing* (pp. 20–39). Reston, VA: Council for Exceptional Children.
Haughton, E. (1980). Practicing practices: Learning by activity. *Journal of Precision Teaching, 1*(1), 29–30.
Heath-Legg, A. J. (1992). *The use of keyboard familiarization and basic word processing in a first-grade spelling program.* Unpublished M.S. practicum, Nova University, Fort Lauderdale, FL. (ERIC Document Reproduction Service No. ED348663)
Heward, W. L. (1994). Three "low-tech" strategies for increasing the frequency of active student response during group instruction. In R. Gardner III, D. M. Sainaito, J. O. Cooper, T. E. Heron, W. L. Heward, J. W. Eshleman, & T. A. Grossi (Eds.), *Behavior analysis in education* (pp. 283–320). Pacific Grove, CA: Brooks/Cole.
Hirsch, E. D. (1996). *The schools we need and why we don't have them.* New York: Doubleday.
Ivarie, J. J. (1986). Effects of proficiency rates on later performance of a recall and writing behavior. *Remedial and Special Education, 7*(5), 25–30.
Johnson, K. R., & Layng, T. V. J. (1992). Breaking the structuralist barrier: Literacy, numeracy with fluency. *American Psychologist, 47,* 1475–1490.
Judd, W. A., & Glaser, R. (1969). Response latency as a function of training method, information level, acquisition, and overlearning. *Journal of Educational Psychology Monograph, 60*(4 Pt. 2), 1–30.

Kay, R. H. (1992). Understanding gender differences in computer attitudes. *Journal of Research on Computing in Education, 25,* 159–171.

Keller, F. S. (1968). “Good-bye, teacher. . . .” *Journal of Applied Behavior Analysis, 1,* 78–89.

Kelly, R. L. (1996). A functional analysis of the effects of mastery and fluency on maintenance. *Dissertation Abstracts International, 57*(2), 639A. (UMI No. 9616720)

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350–386.

Kohn, A. (1999). *The schools our children deserve.* New York: Houghton Mifflin.

Kubina, R. M., & Morrison, R. S. (2000). Fluency in education. *Behavior and Social Issues, 10,* 83–99.

Kulik, C. L. C., Kulik, J. A., & Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60,* 265–299.

Kulik, J. A., & Kulik, C. L. C. (1987). Review of recent research literature on computer-based instruction. *Contemporary Educational Psychology, 12,* 222–230.

Kulik, J. A., Kulik, C. L. C., & Cohen, P. A. (1979). A meta-analysis of outcome studies of Keller’s Personalized System of Instruction. *American Psychologist, 34,* 307–318.

Lanchantin, S. (1991). *Systematic development of an instructional module on writing topic sentences for fifth grade students*. Master’s thesis, New York Institute of Technology. (ERIC Document Reproduction Service No. ED341072)

Lindsley, O. R. (1996). Is fluency free-operant response-response chaining? *The Behavior Analyst, 19,* 211–224.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 1993, 48,* 1181–1209.

Makepeace, M. D. (1998). The effect of rate vs. number of emissions of an operant behavior on behavior probability. *Dissertation Abstracts International, 59*(4), 1835B. (UMI No. 9828812)

McCarty, A. L. (1999). A comparison of the effects of temporal requirement and response form on retention, equivalence and endurance. *Dissertation Abstracts International, 61*(2), 1066B. (UMI No. 9954044)

McDermott, P. A., & Stegemann, J. H. (1987). *The comparative effects of computer-assisted instruction of motivation and achievement of learning disabled students.* Philadelphia: University of Pennsylvania. (ERIC Document Reproduction Service No. ED309611)

McFarland, R. A. (1928). The role of speed in mental ability. *Psychological Bulletin, 25,* 595–612.

McFarland, R. A. (1930). Experimental study of the relationship between speed and mental ability. *Journal of General Psychology, 3,* 67–96.

Mevarech, Z. R., Silber, O., & Fine, D. (1991). Learning with computers in small groups: Cognitive and affective outcomes. *Journal of Educational Computing Research, 7*(2), 233–243.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy: A reader.* Chicago: Adline.

Nathan, R., & Baron, L. J. (1995). The effects of gender, program type, and content on elementary children’s software preferences. *Journal of Research on Computing in Education, 27*(3), 348–360.

Neisser, U. (1983). The rise and fall of the sensory register. *The Brain and Behavioral Sciences, 6,* 1–54.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the power law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisitions* (pp. 1–55). Hillsdale, NJ: Erlbaum.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Olander, C. P., Collins, D. L., McArthur, B. L., Watts, R. O., & McDade, C. E. (1986). Retention among college students: A comparison of traditional versus Precision Teaching. *Journal of Precision Teaching, 4,* 80–82.

Ormrod, J. E., & Spivey, N. R. (1990). Overlearning and speeded practice in spelling instruction. *Psychological Reports, 67*(2), 365–366.

Otis, A. S. (1962). *Examen Otis-Ottawa pour le recrutement du personnel* [Otis-Ottawa exam for personnel recruitment]. Ottawa, Ontario, Canada: Édition de l'Université d'Ottawa.

Pedego, S. E. (1999). The comparative effects of single and yoked learning channel sets under overlearning and fluency training conditions on the retention of paired associate information. *Dissertation Abstracts International, 60*(4), 1836B. (UMI No. 9928126)

Péladeau, N. (2000). *PracticeMill* (Version 1.2) [Computer software]. Montreal: Provalis Research.

Peterson, L. R. (1965). Paired associate latencies after the last error. *Psychonomic Sciences, 2,* 167–168.

Reynolds, J. H., & Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology, 55,* 297–308.

Rosenshine, B., & Berliner, D. C. (1978). Academic engaged time. *British Journal of Teacher Education, 4*(1), 3–16.

Rosenshine, B., & Stevens, R. (1986). Teaching Functions. In M. Wittrock (Ed.), *Handbook on research in teaching* (pp. 376–391). New York: Macmillan.

Rossi, P., Freeman, H., & Lipsey, M. (1999). *Evaluation: A systematic approach* (6th ed.). Newbury Park, CA: Sage.

Schmidt, F. L. (1996). Statistical significance testing for cumulative knowledge psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Semb, G. B., & Ellis, J. A. (1994). Knowledge taught in school: What is remembered? *Review of Educational Research, 64,* 253–286.

Shirley, M. J., & Pennypacker, H. S. (1994). The effects of performance criteria on learning and retention of spelling words. *Journal of Precision Teaching, 12,* 73–86.

Slavin, R. E. (1987). Mastery learning reconsidered. *Review of Educational Research, 57*(2), 175–213.

Smith, E. S. (1980). The effect of computer-assisted instruction on academic achievement, school daily attendance, and school library usage at Margaret Murray Washington Career Center. *Dissertation Abstracts International, 41*(6), 2431A. (UMI No. 8026342)

Spearman, C. (1914). The theory of two factors. *Psychological Review, 21,* 101–115.

Stallings, J. (1975). Implementation and child effects of teaching practices in Follow Through classrooms. *Monographs of the Society for Research in Child Development, 40* (Serial No. 163).

Stallings, J. (1980). Allocated academic learning time revisited, or beyond time on task. *Educational Researcher, 9*(11), 11–16.

Taymans, J. M. (1985). An investigation of the relationship between response latency during overlearning and retention for learning disabled students engaged in computer-assisted instruction. *Dissertation Abstracts International, 47*(2), 503A. (UMI No. 8608867)

Terrill, R., & Ducharme, R. (1994). Passage secondaire-collégial: Caractéristiques étudiantes et rendement scolaire [High-school-to-college promotion: Student characteristics and academic achievement] (2nd ed.). Montreal, Quebec, Canada: Service Régional d'Admission du Montréal Métropolitain.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26–30.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 25–32.

Thorndike, E. L. (1921). The Psychology of drill in arithmetic: The amount of practice. *Journal of Educational Psychology, 12,* 183–194.

Tinker, M. A. (1931). The significance of speed in test response. *Psychological Review, 38,* 450–454.

Vincent, A. T. (1977). The effects of supplementary computer-assisted instruction on the mathematics achievement and attitude toward mathematics of EMR high school students. *Dissertation Abstracts International, 39*(2), 736A. (UMI No. 7812963)

Vockell, E. L., & Schwartz, E. (1988). *The computer in the classroom.* Santa Cruz, CA: Mitchell.

Weitzman, D. L. (1964). *Effect of tutoring on performance and motivation ratings in secondary school students.* Oakland, CA: Oakland Unified School District. (ERIC Document Reproduction Service No. ED001562)

White, O. R. (1985). Aim * star wars (setting aims that compete): Episode 1. *Journal of Precision Teaching, 6,* 7–13.

White, W. A. T. (1988). A meta-analysis of Direct Instruction in special education. *Education and Treatment of Children, 11*(4), 364–374.

Wittman, T.-K. (1996). The relationship between automatization of multiplication facts and elementary school children's mathematics anxiety. *Dissertation Abstracts International, 57*(1), 108A. (UMI No. 9615505)

Wolf, W., & Stroud, J. (1961). IQ test completion and intelligence. *Journal of Educational Psychology, 52,* 245–253.

Manuscript received December 5, 2002
Revision received May 28, 2003
Accepted May 30, 2003

United States Postal Service

Statement of Ownership, Management, and Circulation

1. Publication Title	2. Publication Number	3. Filing Date
American Educational Research Journal	0 1 8 3 - 0 0	October 1, 2003
4. Issue Frequency: Quarterly	5. Number of Issues Published Annually: Four	6. Annual Subscription Price: $48/$110

7. Complete Mailing Address of Known Office of Publication *(Not printer) (Street, city, county, state, and ZIP+4)*
1230 17th Street, NW
Washington, DC 20036-3078

Contact Person: Linda Dziobek
Telephone: (202) 223-9485

8. Complete Mailing Address of Headquarters or General Business Office of Publisher *(Not printer)*
1230 17th Street, NW
Washington, DC 20036-3078

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor *(Do not leave blank)*

Publisher *(Name and complete mailing address)*
American Educational Research Association
1230 17th Street, NW
Washington, DC 20036-3078

Editors *(Name and complete mailing address)*
SIA: Maenette K. P. Benham
Michigan State University, 419A Erickson Hall
East Lansing, MI 48824

TLHD: Margaret McKeown
Learning Research and Development Center
3939 O'Hara Street, University of Pittsburgh
Pittsburgh, PA 15260

Managing Editor *(Name and complete mailing address)*
Martha Yager
1230 17th Street, NW
Washington, DC 20036-3078

10. Owner *(Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)*

Full Name	Complete Mailing Address
American Educational Research Association	1230 17th St., NW, Washington, DC 20036-3078

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box ☒ None

Full Name	Complete Mailing Address

12. Tax Status *(For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)*
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months *(Publisher must submit explanation of change with this statement)*

PS Form **3526,** October 1999 *(See Instructions on Reverse)*

13. Publication Title: American Educational Research Journal

14. Issue Date for Circulation Data Below: Summer 2003

15. Extent and Nature of Circulation		Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies *(Net press run)*		20,493	19,808
b. Paid and/or Requested Circulation	(1) Paid/Requested Outside-County Mail Subscriptions Stated on Form 3541. *(Include advertiser's proof and exchange copies)*	19,838	19,188
	(2) Paid In-County Subscriptions Stated on Form 3541 *(Include advertiser's proof and exchange copies)*	0	0
	(3) Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Non-USPS Paid Distribution	0	0
	(4) Other Classes Mailed Through the USPS	0	0
c. Total Paid and/or Requested Circulation *[Sum of 15b. (1), (2),(3),and (4)]*		19,838	19,188
d. Free Distribution by Mail *(Samples, compliment ary, and other free)*	(1) Outside-County as Stated on Form 3541	0	0
	(2) In-County as Stated on Form 3541	0	0
	(3) Other Classes Mailed Through the USPS	35	35
e. Free Distribution Outside the Mail *(Carriers or other means)*		0	0
f. Total Free Distribution *(Sum of 15d. and 15e.)*		35	35
g. Total Distribution *(Sum of 15c. and 15f)*		19,873	19,223
h. Copies not Distributed		620	585
i. Total *(Sum of 15g. and h.)*		20,493	19,808
j. Percent Paid and/or Requested Circulation *(15c. divided by 15g. times 100)*		99.8%	99.8%

16. Publicati-on of Statement of Ownership
☒ Publication required. Will be printed in the Fall 2003 issue of this publication. ☐ Publication not required.

17. Signature and Title of Editor, Publisher, Business Manager, or Owner
Linda Dziobek Director of Publications

Date: October 1, 2003

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).