All in all, requiring students to participate in research appears to confer modest benefits (or at least does not appear to cause harm). Most students derive some educational benefit from participating in research, report either positive or neutral reactions to their research experience, and would still choose to participate in research even when presented with an alternative that would require equal time and effort. However, if psychology departments are going to continue justifying required research participation in terms of educational benefit, perhaps researchers should strive to improve students' research experiences. Ideally, students would report more than a modest educational benefit, and most students would report feeling positive rather than merely neutral about their research participation.

References

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073.

Bowman, L. L., & Waite, B. M. (2003). Volunteering in research: Student satisfaction and educational benefits. *Teaching of Psychology, 30,* 102–106.

Britton, B. K. (1979). Ethical and educational aspects of participating as a subject in psychology experiments. *Teaching of Psychology, 6,* 195–198.

Britton, B. K. (1987). Comment on Coulter and on Daniel. *American Psychologist, 42,* 268–269.

Britton, B. K., Richardson, D., Smith, S. S., & Hamilton, T. (1983). Ethical aspects of participating in psychology experiments: Effects of anonymity on evaluation, and complaints of distressed subjects. *Teaching of Psychology, 10,* 146–149.

Coulter, X. (1986). Academic value of research participation by undergraduates. *American Psychologist, 41,* 317.

Ferrari, J. R., & McGowan, S. (2002). Using exam bonus points as incentive for research participation. *Teaching of Psychology, 29,* 29–32.

Korn, J. H. (1988). Students' roles, rights, and responsibilities as research participants. *Teaching of Psychology, 15,* 74–78.

Landrum, R. E., & Chastain, G. D. (1995). Experiment spot-checks: A method for assessing the educational value of undergraduate participation in research. *IRB: A Review of Human Subjects Research, 17,* 4–6.

Leak, G. K. (1981). Student perception of coercion and value from participation in psychology research. *Teaching of Psychology, 8,* 147–149.

Nimmer, J. G., & Handelsman, M. M. (1992). Effects of subject pool policy on student attitudes toward psychology and psychological research. *Teaching of Psychology, 19,* 141–144.

Padilla-Walker, L. M., Zamboanga, B. L., Thompson, R. A., & Schmersal, L. A. (2005). Extra credit as incentive for voluntary research participation. *Teaching of Psychology, 32,* 150–153.

Porter, J. P. (1991). The federal policy for the protection of human subjects. *IRB: A Review of Human Subjects Research, 13,* 8–9.

Reatig, N. (1982). Policy guidelines and peer review at the National Institute of Mental Health and National Institute of Child Health and Human Development. *Advances in Law and Child Development, 1,* 97–151.

Rosell, M. C., Beck, D. M., Luther, K. E., Goedert, K. M., Shore, W. J., & Anderson, D. D. (2005). The pedagogical value of experimental participation paired with course content. *Teaching of Psychology, 32,* 95–99.

Sieber, J. E., & Saks, M. J. (1989). A census of subject pool characteristics and policies. *American Psychologist, 44,* 1053–1061.

Waite, B. M., & Bowman, L. L. (1999). Research participation among general psychology students at a metropolitan comprehensive public university. In G. D. Chastain & R. E. Landrum (Eds.), *Protecting human subjects: Departmental subject pools and institutional review boards* (pp. 69–85). Washington, DC: American Psychological Association.

Note

Send correspondence to David Trafimow, New Mexico State University, Department of Psychology, MSC 3452, PO Box 30001, Las Cruces, NM 88003–8001; e-mail: trafimow@crl.nmsu.edu.

Encouraging Distributed Study: A Classroom Experiment on the Spacing Effect

William R. Balch
The Pennsylvania State University, Altoona

*Two introductory psychology classes (*N *= 145) participated in a counterbalanced classroom experiment that demonstrated the spacing effect and, by analogy, the benefits of distributed study. After hearing words presented twice in either a massed or distributed manner, participants recalled the words and scored their recall protocols, reliably remembering more distributed than massed words. Posttest scores on a multiple-choice quiz covering points illustrated by the experiment averaged about twice the comparable pretest scores, indicating the effectiveness of the exercise in conveying content. Students' subjective ratings suggested that the experiment helped convince them of the benefits of distributed study.*

The spacing effect occurs when distributed study results in better memory than massed study does (e.g., Dempster, 1988). This effect holds true for both the spacing of items in a single list (e.g., Underwood, 1970) and the spacing of practice trials (e.g., Baddeley & Longman, 1978). Moreover, the effect applies to both short items like words (e.g., Verkoeijen, Rikers, & Schmidt, 2004) as well as more complex material like textbook passages (e.g., Kraft & Jenkins, 1981) and classroom lessons (Seabrook, Brown, & Solity, 2005).

By frequent testing, an instructor can effectively help students distribute their study and improve their grades (Fulkerson & Martin, 1981). A complementary method might be to use exercises illustrating the benefits of distributed study to motivate students to adopt this strategy by choice. I present a classroom experiment demonstrating the spacing effect.

Method

Participants

One hundred forty-five undergraduates (84 women and 61 men) at The Pennsylvania State University, Altoona, distributed among two introductory psychology classes, participated during their first class of the Fall 2005 semester.

Materials

I selected 16 experimental words from a random sample of 120 common nouns (Zechmeister & Nyberg, 1982, p. 350) originally from the Spreen and Schulz (1966) norms. The experimental words were *message, basket, fashion, justice, artist, supper, ticket, remark* (Subset A); and *cousin, leather, witness, pattern, bottle, empire, giant, habit* (Subset B). Every word contained two syllables and each word in Subset A was approximately matched—on the dimension of concreteness—with a word in Subset B.

Procedure

First, I instructed students that I would read a list of words for them to remember and that most of the words would occur twice somewhere on the list. After the last word, they would count backwards by threes from a two-digit number that I gave them, at a rate of 1 number per second. Then, when I said "Recall," they would write down, in any order, as many words as they could remember.

Reading one word every 3 sec, I presented a list of 36 words to each class. In the first four positions were two-syllable buffer words—*vessel, household, household, and tower,* in that order—to control for the primacy effect. In the remaining 32 positions were two occurrences each of the 16 experimental words. I counterbalanced across the two classes the assignment of the aforementioned word subsets (A or B) to spacing conditions (massed or distributed).

To generate the order of experimental words, I assigned one number to each of the massed words (twice-occurring words occupying two consecutive positions) and two numbers to each of the distributed words (twice-occurring words with at least one other intervening word). Drawing the numbers randomly (with the constraints that neither the numbers of two different massed words nor the numbers of the same distributed word occurred consecutively) resulted in 0, or an average of 12, intervening positions between the occurrences of each massed or distributed word, respectively.

After presenting the words, pacing students through 18 sec of backward counting with a hand motion every second and giving them the signal to recall, I allowed 2 min for the recall task. Then, referring to a transparency of the massed and distributed words projected on a screen, students scored their recall protocols and indicated—by a show of hands—how many remembered more massed than distributed words, more distributed than massed words, or the same number of each word type.

In the debriefing, I identified the experimental hypothesis (people recall more distributed than massed words), the independent variable (massed or distributed words), the dependent variable (number of words recalled), and control procedures. These control procedures included the use of buffer words to eliminate the primacy effect, the backward counting task to eliminate the recency effect, and counterbalancing across classes to control for differences in the memorability of specific words.

To evaluate objectively the effectiveness of the experiment as a teaching tool, all students received a content quiz on the same class day as the experiment. This quiz—taken as either a pretest (Class 1) or posttest (Class 2)—consisted of 10 multiple-choice questions. Seven questions (on massed practice, the relation of massed practice to cramming, distributed practice, the use of buffer words, primacy, recency, and free recall) related specifically to the spacing-effect demonstration and three questions (on recognizing novel examples of an independent variable, a dependent variable, and a counterbalanced variable) related to experimental method in general. As a subjective evaluation, students rated the exercise on convincingness (how much it helped convince them of the benefits of distributing their study over time) and enjoyment.

Results

Spacing Effect

A one-way, within-participants ANOVA combined across classes revealed that students recalled significantly more distributed (47.8%) than massed words (34.5%), $F(1, 144) = 44.74$, $p < .001$, $\eta^2 = .24$ (see Table 1). This effect was significant for both women, $t(83) = 5.98$, $p < .001$, and men, $t(60) = 3.29$, $p < .005$.

Objective Evaluation

Students averaged 38.3% (Class 1) on the 10-question pretest and 77.9% (Class 2) on the identical posttest (see Table 1). A one-way, between-groups ANOVA revealed that this difference was significant, $F(1, 143) = 308.04$, $p < .001$, $\eta^2 = .68$. On the seven spacing-effect questions, students averaged 40.7% on the pretest versus 84.1% on the posttest, $t(143) = 16.58$, $p < .001$. The analogous scores on the three experimental method questions were 32.9% versus 67.1%, respectively, $t(143) = 6.33$, $p < .001$. Better posttest than pretest performances suggest that the exercise effectively presented concepts related to research methodology in general as well as to material specifically related to the demonstration.

Subjective Evaluation

Averaged across classes, students rated the exercise 8.1 on convincingness and 7.2 on enjoyment, based on a scale from 0 (*lowest*) to 10 (*highest*; see Table 1).

Discussion

Although researchers are still trying to determine the best theory of the spacing effect (e.g., Toppino & Bloom, 2002), the robustness, reliability, and broad scope of this effect is clear (Dempster, 1988). Moreover, teachers have several likely means of improving their students' academic performance through distributed study, including frequent testing (Fulkerson & Martin, 1981) and exercises illustrating the benefits of distributed study. Regarding the latter approach, I have extended the strategy of Weseley (1999), who assigned students the task of evaluating the effectiveness of distributed study (and other methods) in terms of evidence from the research literature. The rationale for my technique of demonstrating the spacing effect directly to students was to make the advantage of distributed study as real as possible to them.

By having students score their own recall protocols, I could illustrate the spacing effect immediately (albeit roughly) through a show of hands. About two thirds of the students personally showed the effect, remembering more distributed than massed words. On subsequent checking, I found that about 90% of the students had scored themselves correctly and that only 3% had made errors that changed their conclusions about whether they had obtained the effect. Comparing the average convincingness ratings of students who had shown the effect (8.14) with the ratings of those who had not (7.94), I found no significant difference, $t(143) < 1$, $p > .10$. Apparently, the students based their judgments on the overall results of their class.

My students' relatively high convincingness ratings (M = 8.1 out of 10) directly after the demonstration, although encouraging, did not indicate what effect the demonstration may actually have had on students' subsequent studying. To address this problem, I recently asked 79 additional introductory psychology students to rate both their pre- and postdemonstration study habits 11 weeks after participating in the spacing-effect demonstration during the Spring 2006 semester. They made a vertical mark on each of two horizontal-line scales labeled massed study at one end and distributed study at the other; then I converted their marks to a numerical scale from 0 (*massed*) to 10 (*distributed*). Their postdemonstration distributed-study rating (4.3) was significantly higher than their predemonstration rating (2.6), $F(1, 78) = 137.07$, $p < .001$, $\eta^2 = .64$. Because these ratings were subjective, retrospective, and possibly biased due to my asking for both ratings at the same time, they might not accurately represent students' actual patterns of study. Yet to the extent that the ratings were informative, they suggest that the students tended to cram both before and after the demonstration but that the demonstration may have reduced their cramming slightly.

Instructors who try this demonstration should consider counterbalancing massed and distributed words across word subsets in two different classes and sharing the results with each class. In addition, instructors might discuss the relation between the spacing effect and the practice of cramming for tests. Finally, they might ask their students to suggest future experiments that address the following issues: testing the spacing effect on performance in a real college course (rather than on word recall), objectively assessing the influence of spacing effect demonstrations on students' actual study habits, comparing the effectiveness of such demonstrations with that of a lecture on the same topic, and examining the possible effects of age and other participant variables on the distribution of study. In this way, instructors might stimulate a postdemonstration discussion integrating the spacing effect and general research methodology.

Table 1. Word Recall, Content Scores, and Subjective Ratings in Two Different Classes

	Word Recall[a]		Content Score[b]		Subjective Rating[c]	
Sample	Distributed	Massed	Pretest[d]	Posttest[d]	Convincingness	Enjoyment
Class 1[e]						
M	48.2	37.2	38.3		8.1	7.1
SD	18.2	18.2	17.5		2.1	2.3
Class 2[f]						
M	47.5	31.6		77.9	8.0	7.2
SD	19.1	14.5		19.9	1.8	2.0

Note. $N = 145$.
[a]Percentage of eight distributed versus eight massed words, varied within participants. [b]Percentage of 10 multiple-choice questions correctly answered on a content quiz. [c]Based on a scale from 0 (*lowest*) to 10 (*highest*). [d]Each class received either a pretest or a posttest. [e]$n = 75$, with Subset A words distributed and Subset B words massed. [f]$n = 70$, with Subset A words massed and Subset B words distributed.

References

Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics, 21,* 627–635.

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

Fulkerson, F. E., & Martin, G. (1981). Effects of exam frequency on student performance, evaluations of instructor, and test anxiety. *Teaching of Psychology, 8,* 90–93.

Kraft, R. N., & Jenkins, J. J. (1981). The lag effect with aurally presented passages. *Bulletin of the Psychonomic Society, 17,* 132–134.

Seabrook, R., Brown, G. D., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Journal of Applied Psychology, 19,* 107–122.

Spreen, O., & Schulz, R. W. (1966). Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior, 5,* 459–468.

Toppino, T. C., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28,* 437–444.

Underwood, B. J. (1970). A breakdown of the total-time law in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 9,* 573–580.

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Schmidt, H. G. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30,* 796–800.

Weseley, A. J. (1999). Applying the principles of learning and memory to students' lives. In L. T. Benjamin, B. F. Nodine, R. M. Ernst, & C. Blair-Broeker (Eds.), *Activities handbook for the teaching of psychology* (Vol. 4, pp. 183–185). Washington, DC: American Psychological Association.

Zechmeister, E. B., & Nyberg, S. E. (1982). *Human memory: An introduction to research and theory.* Monterey, CA: Brooks/Cole.

Notes

1. I thank Cynthia Stewart and Althea Eaton for their assistance in conducting, and in scoring data from, the classroom experiment reported in this study.
2. Send correspondence to William R. Balch; Psychology Department; The Pennsylvania State University, Altoona; 3000 Ivyside Park; Altoona, PA 16601; e-mail: wrb3@psu.edu.

The Elusive Definition of Outliers in Introductory Statistics Textbooks for Behavioral Sciences

Thomas P. Hogan and Kimberly Evalenko
University of Scranton

We examined treatment of outliers in 40 introductory statistics textbooks for behavioral science. The majority of texts (75%) included treatment of outliers. Other than agreement in defining outlier as an extreme data point, there was great diversity in treatment. Of the 30 texts including outliers, 11 presented both the univariate and bivariate cases, 15 treated only the univariate case, and 4 only the bivariate case. The texts included 7 different operational definitions of outliers; 16 texts provided no operational definition. A supplementary analysis of outliers in 3 statistical software packages showed more consistency, although not complete agreement. We offer recommendations for treatment of outliers in introductory statistics courses for behavioral science.

Following Tukey's (1970, 1977) pioneering work on exploratory data analysis, the concept of an outlier has become well established in the statistical literature. Research articles often refer to the occurrence and disposition of outliers in data analysis. Most introductory statistics textbooks for the behavioral sciences, as shown in this study, include reference to outliers. In addition, widely used statistical software packages include procedures for identifying outliers.

Many topics in the introductory statistics course (e.g., standard deviations, Pearson correlations) have standardized definitions; textbooks differ mainly in pedagogical techniques (e.g., number and types of examples used) rather than in substance. Formulas for calculating most routine statistics are highly standardized in different textbooks. By contrast, beyond the common definition of an outlier as an aberrant or very unusual data point, we have noticed considerable variety in how textbooks treat this concept. We investigated how introductory textbooks on statistics in the behavioral sciences cover outliers. We supplemented the textbook analysis by analyzing outliers in three statistical software packages.

For purposes of reference, we constructed Figure 1 showing the principal terms used in Tukey's (1970) original treatment of outliers. The figure applies Tukey's terms to a normal distribution of *T* scores ($\mu = 50$, $\sigma = 10$). To keep the figure size manageable, we included only even-numbered scores. Where important reference values in *T* scores or σ units occur at odd-numbered scores, we placed them at the nearest even-numbered score. The crucial parts of Tukey's scheme are (a) use of the interquartile range (IQR) as the basic unit (rather than σ units) for measuring distances; (b) a multiplier of 1.5 applied to the IQR; and (c) a two-tiered system, going $1.5 \times$ IQR, then $3 \times$ IQR beyond the first and third quartiles. Tukey's original presentations did not include the term outlier. Rather, he labeled data points in the first tier as *outside values* and data points in the second tier as *far out values.* Today, use of the term outlier is nearly universal.

Method

We examined 40 introductory statistics textbooks (see Appendix) using several inclusion criteria. The text had to be aimed at the social and behavioral sciences; we did not include mathematical statistics books. The text had to be intro-