Effectiveness of Computer-Assisted Instruction in Statistics : A Meta-Analysis

Giovanni W. Sosa, Dale E. Berger, Amanda T. Saw and Justin C. Mary

REVIEW OF EDUCATIONAL RESEARCH 2011 81: 97 originally published online 12 August 2010
DOI: 10.3102/0034654310378174

Review of Educational Research
March 2011, Vol. 81, No. 1, pp. 97–128
DOI: 10.3102/0034654310378174

Effectiveness of Computer-Assisted Instruction in Statistics: A Meta-Analysis

Giovanni W. Sosa, Dale E. Berger, Amanda T. Saw, and Justin C. Mary
Claremont Graduate University

*Although previous meta-analyses have documented the efficacy of computer-assisted statistics instruction, the current study examined a range of specific features that presumably influence its effectiveness, such as the level of learner engagement, learner control, and the nature of feedback. In 45 experimental studies with a control condition, computer-assisted statistics instruction provided a meaningful average performance advantage (*d = *0.33). Because of great methodological heterogeneity among the studies, the authors employed a conservative but appropriate mixed effects model to examine potential moderator effects. The authors' analyses revealed three statistically significant findings. Larger effects were reported in studies in which treatment groups received more instructional time than control groups, in studies that recruited graduate students as participants, and in studies employing an embedded assessment. A newly developed second order standardized mean effect size,* d_{diff}, *reveals that additional study characteristics may serve as meaningful moderators. Tight experimental control is needed to assess the importance of specific instructional features in computer-assisted statistics instruction.*

Keywords: classroom research, cognitive development, cognitive processes, computers and learning, d_{diff}, meta-analysis, statistics.

Computer-based educational tools have become pervasive in statistics education (Hsu, 2003). The Multimedia Educational Repository for Learning and Online Teaching (MERLOT; www.merlot.org) now includes more than 300 distinct resources, identified by a search for *statistics* within Learning Materials. More broadly, a Google search of the term *statistics tutorial* generates well more than 17,000 internet sites, reflecting a huge body of web-based statistics resources.

Computer-assisted statistics instruction has many potential benefits. First, computerized exercises provide students with additional practice that can reinforce their understanding of the material (Gonzalez & Birch, 2000). Computer-assisted instruction may also allow students to exercise control over the pace at which information is presented (Frederickson, Reed, & Clifford, 2005; Gonzalez & Birch, 2000). Having some control over the learning process has the potential of making instructional tools more relevant to students; moreover, to the extent that

students feel that they have control over the learning process, learner control may also be conducive to increasing students' engagement and expectations for successful understanding of the material (Milheim & Martin, 1991). Practice with tasks and exercises that are relevant to students' everyday life or are meaningful in a given discipline can enhance students' understanding of statistical concepts (Larreamendy-Joerns, Leinhardt, & Correador, 2005; Schumm et al., 2002). This authenticity provides a context for understanding abstract statistical concepts, encouraging interactivity, and facilitating students' ability to apply their knowledge. Computer-based statistics tools that offer students control over how material is presented have the potential of increasing student engagement with course content (Larreamendy-Joerns et al., 2005).

Another benefit of computer-based statistics tools is the potential for providing effective feedback. Feedback that confronts misconceptions and promotes mindful processing of information can facilitate learning (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Moreno, 2004; Moreno & Mayer, 2005; Timmerman & Kruepke, 2006). Feedback can be corrective (indicate whether the user is accurate), be explanative (offer more in-depth reasoning), reinforce a correct response, or support self-regulated learning. Feedback may be immediate or delayed, specific or more general. Feedback has been found to be most powerful when it addresses misconceptions rather than merely dealing with a lack of knowledge and when the exercise or task complexity is low (Hattie & Timperley, 2007). For instance, Moreno and Mayer (2005) found that in an interactive game setting, guidance in the form of explanatory feedback improved learning and reduced misconceptions; however, merely having the users explain their responses was less effective.

Hattie and Timperley (2007) concluded that feedback is most effective when goals are clearly defined and feedback is highly related to attaining the goals in question. Feedback is further enhanced when it indicates success with current activities, includes information on what is understood and what is not, and guides the learner through activities and strategies needed to reach specified goals. Hence, effective feedback not only is based on monitoring progress toward the specified goal but also promotes effective learning strategies.

Learning goals should include promoting conceptual knowledge (i.e., statistical thinking and reasoning) in addition to procedural learning and using assessments to improve instruction, not merely to rank students (Garfield, 1994; Garfield & Chance, 2000). Garfield and Chance (2000) suggested multiple forms of assessment, including minute papers, projects, and portfolios in addition to traditional forms such as quizzes and exams, to measure deeper conceptual understanding and to help students achieve learning goals. In addition, assessment that is incorporated into the learning activity may offer instructors the opportunity to evaluate student learning during the learning process itself and can offer students and faculty the feedback necessary to improve both student learning and instructional techniques (Kennedy, Brown, Draney, & Wilson, 2005). These issues related to goals and feedback highlight the importance of assessment methods, particularly in the domain of computer-assisted instruction where assessment can be built into the technological tools.

Computer-based statistics tools can be designed to offer the user immediate and response-specific feedback (Larreamendy-Joerns et al., 2005; Morris, 2001)—something that is virtually impossible in large classroom settings. As with learner

control, the timing of feedback is inherently linked to a tool's overall interactivity. Rapid feedback has proven helpful in making course content more meaningful to students not only by increasing motivation but also by allowing students to confront misconceptions quickly (Aberson, Berger, Healy, & Romero, 2003; Svinicki, 1999). Effective tools offer learners response-specific feedback that is tailored to specific learning needs, and they provide learners with an appropriate amount of meaningful information.

Compared to classroom instruction, many computer-based tools require greater active engagement from students. The computational power afforded by such tools allows students to devote more time and energy toward mastering the conceptual, rather than procedural, facets of statistical concepts. Students may be asked to actively manipulate their learning environments by, for instance, altering parameters such as sample size, effect size, or *p* value. Computer-based tools are able to offer more examples and a wider array of examples than a student would normally be exposed to in the classroom. Interactivity and engagement in computer-based instruction can enhance statistical understanding by promoting deeper processing (Chi, 2009) and reducing cognitive load (Moreno & Mayer, 2007). By promoting individualized active learning in a structured environment, interactive technology can enable students to organize information so that it can be connected to their prior knowledge or existing schemata.

The many potential benefits that computer-based tools offer have spurred the development of an assortment of resources; however, abundance and diversity of computer-based resources challenge educators who wish to select tools that will be most beneficial. To address this issue, MERLOT established peer-review procedures meant to describe the strengths and weaknesses of individual computer-based learning tools and to identify especially effective resources (Cafolla, 2006). Specifically, MERLOT reviewers evaluate educational resources based on three broad dimensions: (a) quality of content, (b) potential effectiveness as a teaching and learning tool, and (c) ease of use.

A peer-review process such as the one employed by MERLOT has the potential to help educators and students sort through the many existing tools to identify those best suited for their needs. However, subjective ratings are not equivalent to an empirical outcome evaluation. For instance, Jolicoeur and Berger (1986) found that experts showed surprisingly low interrater reliability in their judgments of the quality of educational software. In a subsequent study, Jolicoeur and Berger (1988a, 1988b) used random assignment with control conditions to test the effectiveness of six computer-based educational programs. At the conclusion of the study, both teachers and students judged the educational effectiveness of each software package. These subjective measures of perceived educational effectiveness were poorly correlated with actual educational effectiveness. Thus, although subjective ratings may be helpful in describing and classifying an assortment of computer-based tools, empirical evaluations are necessary to establish actual effectiveness.

A meta-analysis reported by the U.S. Department of Education, Office of Planning, Evaluation, and Policy Development (2009) suggests that in general online instruction can be more beneficial than traditional face-to-face instruction for both K–12 students and older learners across a range of learning domains. Based on 51 effect sizes, this meta-analysis found the overall mean effect size *d*,

the number of standard deviations the treatment mean exceeds the control mean, to be 0.24, favoring the use of online learning. Meta-analyses by James Kulik and his colleagues (C. Kulik & Kulik, 1991; C. Kulik, Kulik, & Shwalb, 1986; J. Kulik, 2002; J. Kulik & Kulik, 1987) documented the effectiveness of computer-based instruction specifically in mathematics over the span of two decades. Their findings consistently indicate that students using computer-based tools outperform students receiving only lecture-based instruction. J. Kulik (2002) estimated that across 200 studies the overall mean effect size *d* is 0.30. If the underlying distributions are normally distributed, an effect size this large means that the performance of the average student receiving computer-based instruction in mathematics exceeds the performance of 62% of students receiving lecture-based instruction. Although such an effect size may be considered small according to Cohen's (1977) conventions, Slavin (1990) argued that in an educational context an effect size of 0.25 or greater has practical significance. Thus, Kulik's findings serve to confirm that computer-based tools can effectively contribute to learning, at least in the field of mathematics.

The evidence for the effectiveness of computer-based tools in teaching statistics appears to be at least as strong as in the broader field of mathematics. Hsu (2003) identified 25 empirical studies published between 1986 and 2002 that examined the effectiveness of computer-assisted instruction in statistics. She found an impressive average effect size of 0.43. More importantly, however, her study revealed potential moderating variables. She found that computer-based tools described as drill and practice resources (tools that present users with a set of test questions or practice problems that need to be completed in a specified order) produced the largest effects. She also found that instructor-made programs were more effective than commercially available programs. Similarly, Schenker (2007) identified 46 empirical studies examining the effectiveness of technology used in statistics instruction. Schenker obtained a mean effect size of 0.24. In examining the individual effect sizes, he found drill and practice programs to be the only technology type significantly associated with the apparent advantage for computer-assisted instruction. Significantly larger effects were also found among studies utilizing control groups that received no instruction and studies that examined the effectiveness of computer-assisted instruction in the context of math courses. On the other hand, Schenker found that online learning tools yielded lower achievement outcomes when compared to the other technology types.

Unlike Hsu (2003), Schenker (2007) assigned some studies to multiple technology categories (i.e., a given study could be designated as both a drill and practice tool and a tutorial), confounding the comparisons across technology types. A related issue is the overlap among identified technology types. Hsu and Schenker identified seven and eight different technology types, respectively, with varying degrees of overlap in the conceptualization of these tools. For instance, Schenker made a distinction between online learning and technology-enhanced lectures; although the former are described as online learning environments that offer students access to materials and web-based resources such as videos, images, and sounds, the latter are described as offering students access to similar multimedia resources. Redundancy between identified technology types complicates assessment of the effects of different types of tools. An alternative is to use broader

categories that reduce overlap and provide a larger number of studies within categories. For example, most technology tools for teaching statistics can be categorized as stand-alone tutorials, number crunchers, or communication tools. Stand-alone tools are characterized by tutorials or exercises that students can complete on their own with little or no support from an instructor (e.g., http://wise.cgu.edu). Number crunchers are computational aids, such as SPSS or Minitab. Communication tools include class support systems such as Blackboard or Sakai. Although some evidence exists regarding the benefits of specific technological tools, more work is needed to elucidate our understanding of this matter.

Another limitation of previous meta-analytic work is that learner-centered aspects, such as student engagement, student control, and the nature of feedback, were not explored. The literature cited earlier indicates that such features are critical to facilitating the learning process and promoting the development of critical thinking skills (Lovett & Greenhouse, 2000; Mills, 2002; Svinicki, 1999). Little empirical work has examined the direct impact that such features of computer-assisted instruction have on the learning of statistical concepts.

The Current Study

The meta-analyses by Hsu (2003) and Schenker (2007) clearly established that computer-based tools can be effective in statistics instruction. However, these reviews did not focus attention on aspects of computer-based tools that are most closely associated with learning and achievement outcomes. The current study extends previous work by giving attention to additional attributes that could account for differences in effectiveness, such as different technology types, student engagement, student control over the learning process, and the nature of feedback.

Method

Search Procedure

The goal of the literature search was to identify all studies that reported an empirical outcome evaluation of computer-assisted instruction in statistics in comparison to a control condition. We searched the following databases: (a) Educational Resources Information Center, (b) ProQuest Digital Dissertations, (c) Ingenta, (d) OmniFile, and (e) PsycINFO. Our search phrases included various terms coupled with the keyword *statistics*. These terms included *computer*, *computer-based*, *computer-assisted*, *distance learning*, *distance education*, *web instruction*, *tutorial*, *simulation*, *technology*, *internet*, *applet*, and *software*. In addition, the keyword *effectiveness* was added to each of the aforementioned phrases; this last procedure was repeated with the keywords *evaluation*, *assessment*, and *performance*. That is, we searched every combination of the term *statistics* with one of the technology words (e.g., *computer*) and one of the outcome terms (e.g., *evaluation*). For instance, our initial search was *statistics and computer and evaluation*. We also searched the *Journal of Statistics Education* and the studies used by Hsu (2003) and Schenker (2007) in their meta-analyses. We first identified potential studies by reading through all titles and abstracts resulting from our searches. After a potential study was identified, the full body of the article was examined to determine whether the study met our inclusion criteria.

Studies were included only if they met all of the following criteria: (a) they assessed the effectiveness of computer-based instructional tools in statistics, (b) they evaluated effectiveness based on objective performance or learning measures (i.e., test scores), (c) they provided comparison data from a lecture-based control group receiving instruction in statistics (thus, simple pre–post studies, studies in which all groups were exposed to computer-based tools, and studies in which the control group did not receive instruction were excluded), (d) the control group did not receive any form of computer-assisted instruction, (e) they provided sufficient information to calculate an effect size, and (f) they provided enough information for the authors to be able to rate the studies with regard to specific learner-centered features. One article (Athey, 1987) described two unique studies with different participants—these studies were coded individually. Other articles with multiple studies employed the same treatment and/or control participants across studies. In such cases, data were pooled to generate a single effect size estimate. This resulted in a total of 45 unique studies that met our inclusion criteria.

Coding

Bibliographic information. Bibliographic information was coded as follows: (a) study ID number, (b) complete American Psychological Association (APA) citation, (c) type of publication report (e.g., journal article, conference presentation, or doctoral dissertation), and (d) the year of publication.

Sample descriptors. Descriptors of the sample were coded as follows: (a) academic level of participants (high school, college, or graduate students), (b) whether participants assigned to the computer-assisted instruction group possessed any formal training in statistics prior to the intervention, (c) the treatment group sample size at the start and end of the study, and (d) the control group sample size at the start and end of the study.

Research design descriptors. Several design features were coded dichotomously. First, we coded the unit of assignment to conditions. In some cases, participants were assigned individually to treatment groups; in other cases, entire classrooms were assigned to treatment conditions. Second, we coded whether assignment was random or not random. Third, we coded whether the equivalence of the groups on the learning or performance measure of interest was assessed at pretest. Last, we documented whether the comparison between groups used statistical control of any potential covariates (e.g., GPA or mathematics ability).

Treatment descriptors. We coded multiple features of the computer-based technology and its implementation. One feature we coded was whether the computer-based tool was used to replace or only supplement classroom-based instruction, that is, whether the tool replaced activities of the instructor or whether it was used to augment lecture-based instruction. Another feature we coded was the number of sessions during which participants were exposed to the computer-based tool; in some studies participants utilized a tool on a single occasion, whereas in other studies they used a tool on multiple occasions. In addition, we coded whether the tool was implemented in the context of a distance education course, and we coded

whether the tool was produced commercially (e.g., SPSS) or developed by the researcher or instructor.

We also examined the issue of differential exposure to statistics instruction between the treatment and control groups. We found that most studies made a careful effort to ensure that the two groups received an equal amount of instruction by controlling the number of hours or sessions. However, we identified six studies in which it was clear that the computer-instruction group received more overall instruction than did the control group. Last, we also coded for the type of outcome assessment used in studies. Some studies employed an embedded assessment approach in which the target items were incorporated into a broader assessment (e.g., midterm or final exam). Others crafted an assessment tool entirely composed of target items and specifically developed for the study in question. Yet other studies had students in both groups take the final exam that was similar, if not identical, to the one normally used in lecture-only courses. Finally, some studies simply compared the two groups on their final course grade.

As noted earlier, computer-based instruction potentially can enhance individual engagement in processing information by promoting interactivity, allowing students to control the pace of the interaction, requiring thoughtful responses, and providing rapid individualized feedback. These factors have been shown to be important in many learning contexts (Chi, 2009; Halpern & Hakel, 2003; Lovett & Greenhouse, 2000). To assess the extent to which the computer-based instruction provided these features, three authors independently rated specific attributes on a 0–4 scale for each of the 45 studies in our meta-analysis.

With respect to the interactivity of the computer-based tool, we rated three specific attributes. Raters could also abstain from committing to a value if they felt that the study in question lacked sufficient information to make such a determination. First, we rated the extent to which the computer-based tool required thoughtful responses. Next, we rated the extent to which the user exercised control over the pace of the learning process. Last, we rated the extent to which the tool engaged active learning; that is, the degree to which the tool was cognitively engaging. With respect to the nature of feedback provided by the tool, we rated two attributes. First, we rated the extent to which the explanations or feedback stemming from the tool were specific to individual learners' responses (i.e., targeted feedback). Second, we rated the immediacy of the feedback that we perceived to be conducive to learning. In regard to the sources of instructional content, we rated the extent to which the learning process was dependent on instructor involvement, the extent to which the learning process was dependent on interactions among students, and the extent to which the computer-based tool contributed to instruction beyond the instructor or student–student interactions.

In addition, two authors independently rated (a) the complexity of the statistical concepts presented by the computer-based tools, (b) the breadth or range of the statistical topics included in the computer-based instruction, and (c) the extent to which the computer-based tool served as a simulation or an interactive program that allowed students to manipulate information. Complexity was rated on a 0–4 scale. Lower ratings denoted computer-based tools that were calculation based or focused on descriptive information, whereas higher ratings denoted tools that focused on understanding concepts such as hypothesis testing or inferential statistical procedures. Breadth was rated on a 1–4 scale. Lower ratings were assigned to

studies that described computer-assisted instruction that focused on only one or two topics; on the other hand, higher ratings were assigned to studies describing instruction that covered a wide range of topics from basic descriptive statistics to advanced inferential statistics (e.g., hierarchical regression and statistical power). The extent to which a computer-based tool served as a simulation was rated on a 0–2 scale, with higher ratings denoting highly interactive tools that allowed students to manipulate and interact more extensively with simulations.

To assess interrater reliability of subjective ratings, we computed a two-way mixed model intraclass correlation (ICC) based on ratings provided by two or more authors (Shrout & Fleiss, 1979); this analysis was conducted for each of the rated attributes. These ICCs ranged from .80 to 1.0 (average measures) for each evaluated attribute, justifying the use of an average rating across the raters for each attribute to be used in analyses of potential moderating effects.

Last, we examined differences among three broad categories of computer-based tools: (a) stand-alone tools or tutorials, (b) number crunchers, and (c) communication-based instructional tools. Stand-alone tools or tutorials are designed to guide the user through the learning process without the need of an instructor or other third party. Examples of such tools are tutorials offered by the Web Interface for Statistics Education website (http://wise.cgu.edu) and documented in Aberson et al. (2003; Aberson, Berger, Healy, Kyle, & Romero, 2000; Aberson, Berger, Healy, & Romero, 2002). Number crunchers refer to tools that allow users to manipulate or analyze data (e.g., SPSS, STATA). In contrast to stand-alone tools, number crunchers do not offer conceptual or theoretically based instruction and are not designed to take the place of the instructor. Instead, they operate as advanced calculators to provide users with computations and statistical output. Although they may provide some instruction through help menus, they require the user to have the knowledge necessary to select an appropriate statistical technique and the ability to interpret the resulting output. Communication-based instructional tools refer to computer-based resources through which instructors can interact with students (e.g., e-mail, listserver, or a website for a class). For instance, a distance education course may have a website that, in addition to allowing students to download readings and lecture materials, also offers users access to a discussion board or chat sessions. Other examples of communication-based instructional tools are keypads or clickers that allow instructors to pose questions and receive individual feedback from students during a class session. These tools function as electronic media by which the instructor and students can interact. The three authors coding for these characteristics obtained a very high level of agreement (single measure ICC = .98).

Effect size measures. All data needed for calculation of effect sizes were recorded (i.e., means, standard deviations, significance test results). In addition to recording effect size–related information, the first author also recorded a confidence rating for each study's estimated effect size. Specifically, he rated the extent to which each effect size was estimated, based on a 1 (*highly estimated, minimal data provided*) to 5 (*sufficient data provided for computation*) scale. With the exception of one study that offered only significance test results and sample sizes (score of 2), all studies received a score of 3 or higher. All effect sizes were computed as standardized mean difference effect sizes (*d*). Several studies failed to report mean and

standard deviations and instead included only significance test levels and sample sizes. We converted such data into standardized mean difference effect sizes using formulas provided by Lipsey and Wilson (2001).

Standardized mean difference effect sizes are known to be upwardly biased when based on small sample sizes (Lipsey & Wilson, 2001). Therefore, we employed Hedges's (1981) correction for this bias, resulting in what is referred to as the *unbiased* effect size. To compute the unbiased effect size, the following formula was applied to each obtained effect size,

$$ES'_{sm} = [1 - (3 / (4N - 9))]\, ES_{sm},$$

where N represents the total sample size and ES_{sm} represents the unadjusted standardized mean difference effect size. All subsequent descriptions of effect sizes, standard errors, and inverse variance weights use the unbiased effect size.

Analysis Strategy: Homogeneity Analysis Across the Entire Distribution of Effect Sizes

There are two contrasting approaches to estimating population parameters and examining the homogeneity across the distribution of effect sizes: the *fixed effects model* and the *random effects model*. The fixed effects model assumes that each effect size from all studies included in a meta-analysis serves as an estimate of a single population effect, with participant-level sampling error that stems from random differences between participants (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hedges, 1994; Lipsey & Wilson, 2001). In contrast, the random effects model reflects the perspective that the included studies estimate multiple population effects. Thus, a random effects model assumes that each effect size serves as an estimate of a true population effect with error that stems from participant-level sampling error (as in the fixed effects model) as well as from random study-level differences reflecting variations in study settings, procedures, measures, and so on. Although participant-level sampling error has a variance equal to the observed within study variance, error stemming from random study-level differences has a variance equal to a between-studies variance component (i.e., τ^2) that reflects the variance between the multiple population effects (for a more thorough conceptual and computational discussion, see Borenstein et al., 2009).

In a fixed effects model, individual participants serve as the sampling units; in a random effects model, however, individual studies serve as the sampling units. Consequently, in a fixed effects model, generalizability is limited only to other participants who could have been recruited for the specific type of study documented by the meta-analysis (Rosenthal, 1995). In contrast, findings from a random effects approach can be generalized to any study belonging to the same population of studies from which effect sizes were obtained (Hedges & Pigott, 2004). Although improving generalizability, the drawback of the random effects model is a reduction in statistical power, largely the result of including the between-studies variance component, τ^2, because this component is added to each of the variances in the obtained effect sizes (Borenstein et al., 2009; Hedges & Pigott, 2004; Lipsey & Wilson, 2001). Although we report both fixed and random effects parameter estimates, our focus is on findings stemming from the random effects analyses because of the high degree of variability among the included studies along various relevant

study characteristics (e.g., nature of technology, experimental settings, measures). The random effects model also allows us to generalize our findings to the entire population of studies that have been conducted in this area of research.

Results

Table 1 presents summary statistics for both fixed and random effects models. We identified effect sizes from two studies as extreme values ($d = 2.12$ and $d = -0.91$). Rather than eliminating them from the analysis, we Winsorized them (cf. Lipsey & Wilson, 2001) by recoding each value to a more moderate value, namely, the second most extreme value in the corresponding tail of the distribution (1.41 and –0.38, respectively). In line with Lipsey and Wilson's (2001) recommendation, we recalculated each effect's corresponding standard error and inverse variance weight. All findings reported hereafter are based on these Winsorized values. The estimated weighted mean effect size of the 45 studies for the fixed effects model is 0.13, whereas that of the random effects model is 0.33. The difference in the magnitude of the mean effect size between the two analytical approaches reflects the lower weight that a random effects approach places on large sample sizes (Borenstein et al., 2009). The estimated random effects mean effect size estimate is clearly larger than its corresponding fixed effects estimate for these studies; however, the corresponding confidence interval is wider (random effects $CI._{95} = 0.20 \leq \mu \leq 0.46$ vs. fixed effects $CI._{95} = 0.09 \leq \mu \leq 0.17$, respectively).

We further examined the computed mean effect via a coefficient of robustness (CR; Rosenthal, 1995). The CR reflects the homogeneity of the obtained effects (greater homogeneity is associated with a larger CR) and the consistency in the directionality of the result (i.e., the greater the proportion of positive effects in a distribution, the larger the resulting CR). A CR is defined simply as the mean effect size divided by the standard deviation of the effect sizes (Rosenthal, 1995). According to this index, the fixed effects mean effect size shows slightly greater robustness (0.13 / 0.31 = 0.42) than does the random effects mean effect size (0.33 / 0.94 = 0.35). This indicates that the effect sizes stemming from a fixed effects model are slightly more consistently positive, and homogenous, than are those stemming from a random effects analysis. Although no data-based benchmark exists for the identification of meaningful CRs, one suggested threshold is a CR of 0.71 (R. Rosenthal, personal communication, April 18, 2009). Thus, although 75% of our effects are positive and the confidence interval limits are greater than zero, the degree of heterogeneity among the effects is substantial and suggests limited convergence among the findings.

As with any meta-analysis, a legitimate concern is whether the magnitude of our estimated effect size is reflective of all the empirical work, unpublished as well as published, that has compared computer and lecture-based statistics instruction. To address this concern, we computed a fail-safe *N* (Orwin, 1983; Rosenthal, 1979) to determine the number of undiscovered studies with an effect size of zero that would need to exist to reduce the random effects mean effect size from 0.33 to 0.10. We determined that it would take 104 additional studies with an effect size of zero to reduce our estimated effect size to 0.10. Moreover, it would take 254 additional studies with an effect size of zero to reduce the mean effect size from 0.33 to 0.05.

TABLE 1
Statistical summary for performance scores

Number of studies	45
Total number of students	9,639
Median number of students	74
Unweighted mean *d*	0.37
Unweighted median *d*	0.33
Proportion $d > 0.00$	0.76
Minimum unweighted *d*	−0.38
Maximum unweighted *d*	1.41
Fixed effects (weighted)	
Mean *d*	0.13
Standard deviation	0.31
Standard error	0.02
95% confidence interval	0.09 to 0.17
Robustness (*M*/*SD*)	0.42
Random effects (weighted)	
Mean *d*	0.33
Standard deviation	0.94
Standard error	0.06
95% confidence interval	0.20 to 0.46
Robustness (*M*/*SD*)	0.35

Note. These findings reflect the Winsorizing of one outlier on each end of the distribution.

Figure 1 shows the distribution of the 45 un-Winsorized effect sizes. As suggested by the CR reported earlier, homogeneity analysis of the effects revealed that the observed differences between the effect sizes are greater than one would expect based on sampling error alone if the effect sizes were computed from groups of participants who were all selected randomly from the same population, $Q(44) = 216.09$, $p < .0001$. Such heterogeneity indicates that the fixed effects model may not be appropriate; it suggests that weighting effect sizes by participant-level sampling error alone is not enough to fully characterize the amount of observed variability among the effects as estimates of a single population mean effect size (Lipsey & Wilson, 2001). Thus, our findings indicate that all effect sizes do not estimate the same single population mean.

Heterogeneity and Between-Subgroup Analysis

Three different types of models have been used to analyze heterogeneous effect size distributions: (a) fixed effects models, (b) random effects models, and (c) mixed effects models. Under a *fixed effects model*, the analyst continues to assume that the variability among effect sizes is primarily because of random variation associated with participant-level sampling error. Any excess variability underlying the observed heterogeneity is viewed as systematic; that is, it can be accounted for by moderator variables (Lipsey & Wilson, 2001). This approach provides the greatest statistical power (i.e., it is more likely to identify statistically significant moderators).

Stem	Leaf
2.1	2
2.0	
1.9	
1.8	
1.7	
1.6	
1.5	
1.4	1, 1
1.3	
1.2	
1.1	
1.0	
.9	3, 4
.8	8
.7	6, 8, 9, 9
.6	5, 8, 9
.5	2, 6, 9
.4	0, 5
.3	3, 4, 7, 9, 9
.2	1, 2
.1	0, 2, 4, 7, 8
.0	2, 2, 6, 7, 9
–.0	4, 5, 7, 7, 9
–.1	
–.2	5, 5, 7,
–.3	8
–.4	
–.5	
–.6	
–.7	
–.8	
–.9	3

FIGURE 1. *Stem and leaf display of unweighted effect size* d*s for the 45 studies of computer-assisted instruction in statistics. Effect size* d*s reflect Hedges' (1981) correction and were Winsorized for analyses.*

In contrast, the *random effects model* views heterogeneity as stemming exclusively from random differences among studies (e.g., variations in procedures and settings). A meta-analyst adopting a random effects model for a heterogeneous distribution will not proceed to identify systematic differences between effect sizes (i.e., moderator analysis) because the excess variability among effect sizes within this model is attributed solely to random—rather than systematic—differences between studies (Lipsey & Wilson, 2001).

A third approach to the analysis of a heterogeneous distribution is the *mixed effects model*. In a mixed effects model, as in a random effects model, the sampling units are not the individual participants but rather the individual studies identified by the analyst. However, the mixed effects approach assumes that the variance beyond sampling error because of participants can be accounted for by both moderators (as in a fixed effects model) and random study-level differences (as in the random effects approach; Hedges & Pigott, 2004; Lipsey & Wilson, 2001). The benefit of a mixed effects model is that findings can be generalized to any study belonging to the same population of studies from which effect sizes were obtained (Hedges & Pigott, 2004). The drawback to this approach, however, is reduced statistical power compared to a fixed effects model (Hedges & Pigott, 2004; Lipsey & Wilson, 2001).

Given the diversity in the types of applications employed by the identified set of studies, along with the rapid growth of new applications, we anticipate that computer-assisted instruction will continue to encompass a highly diverse range of applications. Because of this diversity among studies, we chose a mixed effects model as the most appropriate method by which to examine the presence of potential moderators. We used a noniterative method of moments approach to estimate the between-studies variance component (Borenstein et al., 2009; Lipsey & Wilson, 2001). Analyses were conducted with SPSS macros developed by Wilson (2005; see http://mason.gmu.edu/~dwilsonb/ma.html).

Moderator Analyses With Categorical Study Descriptors: The Utility of d_{diff}

After finding that only a single study recruited high school students, we dichotomized our index of academic level to reflect studies that included undergraduate ($n = 29$) or graduate ($n = 10$) student participants. In addition, because many studies simply reported that the treatment group used the computer-based tool throughout the academic term without specifying an exact number of sessions, we dichotomized our index of the number of sessions during which participants utilized a computer-assisted tool into one session versus two or more sessions.

Findings stemming from the mixed effects moderator analyses for 24 categorical variables measuring a wide range of study characteristics are summarized in Table 2. In addition to indicating the average weighted standardized mean effect size (d) for each subgroup, Table 2 presents the 95% confidence intervals for each effect. Table 2 also shows p values for the comparisons between subgroups across each of the examined study characteristics. Although a p value serves as an informative index of statistical differences between groups, it clearly confounds two sources of information: sample size and effect size (Borenstein et al., 2009; Wilkinson & the Task Force on Statistical Inference, 1999). Awareness of this limitation has spawned strong recommendations that effect sizes be reported as standard practice (e.g., APA, 2001). Indeed,

effect size indices—unlike *p* values—drive conclusions we make about the strength of the relationship among variables, in both primary and meta-analytic studies. However, despite the recognition that *p* values are inadequate measures of practically significant relationships among variables, and despite the recognition that effect size indices reveal meaningful information about the magnitude of study effects, meta-analyses continue to rely on *p* values to identify moderator variables. Although the meta-analytic approach has helped to reduce an overreliance on *p* values, meta-analysts typically identify subgroup differences (i.e., moderator variables) on the basis of the resulting *p* values rather than effect sizes.

To measure the strength of the relationship between study characteristics and the magnitude of the effects, we also report d_{diff}, a new index that operates as a second-order standardized mean effect size (see Table 2). The computation of this index is identical to the computation of a standardized mean effect size, except that it is based on study-level characteristics. As such, it is defined as,

$$d_{diff} = \frac{\overline{d}_{G1} - \overline{d}_{G2}}{S_{d_{pooled}}} \qquad S_{d_{pooled}} = \sqrt{\frac{S^2_{\overline{d}1}(k_1 - 1) + S^2_{\overline{d}2}(k_2 - 1)}{k_1 + k_2 - 2}}$$

where $\overline{d}_{G1}$ and $\overline{d}_{G2}$ reflect the weighted mean effects for each of the subgroups under comparison, $S_{dpooled}$ represents the pooled variance of the subgroups in question, $S^2_{\overline{d}1}$ and $S^2_{\overline{d}2}$ reflect each subgroup's individual variances, and k_1 and k_2 represent the number of studies per subgroup. The variances ($s^2_{\overline{d}1}$ and $s^2_{\overline{d}2}$) used in computing this index have already been adjusted to reflect the corresponding between-studies variance component, τ^2; thus, this index can easily be used with both fixed and mixed effects models. One benefit of d_{diff} is that it avoids the capitalization of chance that is inherent in conducting multiple comparisons relying on *p* values (Lipsey & Wilson, 2001). Another benefit of this index is that it can be interpreted in much the same way that one would interpret any standardized mean effect size—it characterizes the number of standard deviation units that separate the mean effects of the subgroups under comparison. Thus, d_{diff} can be interpreted as a measure of practically significant differences between the subgroups under comparison. Because the magnitude of this index is not driven by the number of participants or the number of included studies, d_{diff} provides useful alternative information for identifying moderator variables. As is illustrated in Table 2, we report d_{diff} along with a corresponding 95% confidence interval for each of our categorical variables.

More study time. Studies where treatment groups devoted more overall time for instruction compared to their control groups showed an average $d = 0.97$, an effect statistically superior to the average $d = 0.23$ found for studies where treatment and control groups received an equal amount of instruction, $Q(1) = 18.53$, $p < .001$. Assuming normal distributions for student performance within conditions, the confidence interval ($CI_{.95} = 0.66 \leq \mu \leq 1.29$) indicates that it is highly likely that 75% to 90% of the students who received additional time outperformed the average student receiving only the control amount of instruction. The d_{diff} of 1.99 indicates that about 97% of the studies affording more total instructional time to students using the computer-based tools are estimated to yield larger effects than the average study not offering computer-based students such an advantage. Thus, it is clear that students who

TABLE 2

Mixed effects moderator analysis examining the weighted mean effect size ds *by study descriptor*

Study descriptor	*k*	*d*	95% CI	*p*	$\lvert d_{diff} \rvert$	*95%* CI (d_{diff})
Additional time[a,b]				< .001	1.99	1.03 to 2.94
Yes	6	0.97	0.66 to 1.29			
No	39	0.23	0.12 to 0.35			
Academic level[a,b]				.006	1.08	0.32 to 1.83
Undergraduate	29	0.25	0.12 to 0.39			
Graduate	10	0.68	0.41 to 0.94			
Embedded assessment[a,b]				.02	0.92	0.09 to 1.75
Yes	7	0.67	0.36 to 0.99			
No	38	0.26	0.12 to 0.40			
Publication type: Conference paper[a,b]				.09	0.71	−0.24 to 1.65
Yes	5	0.04	−0.31 to 0.40			
No	40	0.37	0.23 to 0.52			
Implementation of tool[b]				.15	0.47	−0.18 to 1.13
Replacement	13	0.17	−0.08 to 0.42			
Supplement	32	0.39	0.24 to 0.54			
Pretest[b]				.17	0.44	−0.21 to 1.09
Yes	13	0.19	−0.06 to 0.43			
No	32	0.39	0.23 to 0.55			
Course grade assessment[a]				.24	0.45	−1.26 to 0.36
Yes	7	0.16	−0.17 to 0.48			
No	38	0.37	0.22 to 0.52			
Publication type: Journal[a,b]				.31	0.33	−0.32 to 0.98
Yes	32	0.37	0.22 to 0.53			
No	13	0.23	−0.01 to 0.46			
Dedicated assessment[a]				.35	0.28	−0.86 to 0.31
Yes	22	0.26	0.06 to 42			
No	23	0.39	0.21 to 0.58			
Control of preexisting differences[a,b]				.36	0.28	−0.34 to 0.90
Yes	15	0.25	0.04 to 0.46			
No	30	0.37	0.22 to 0.53			
Origin of tool[a]				.44	0.23	−0.37 to 0.83
Commercially developed	19	0.39	0.19 to 0.60			
Instructor developed	25	0.28	0.10 to 0.47			
Number of sessions				.54	0.21	−0.44 to 0.87
1	13	0.39	0.13 to 0.65			
2+	30	0.29	0.15 to 0.45			
Stand alone vs. communication[a]				.67	0.17	−0.62 to 0.95
Stand alone	12	0.40	0.16 to 0.63			
Communication	13	0.33	0.12 to 0.53			

(continued)

TABLE 2 (continued)

Study descriptor	*k*	*d*	95% CI	*p*	$\lvert d_{diff} \rvert$	*95%* CI (d_{diff})
Distance learning[a]				.64	0.15	−0.51 to 0.82
Yes	12	0.28	0.03 to 0.53			
No	33	0.35	0.20 to 0.49			
Stand-alone tools[a]				.66	0.15	−0.53 to 0.83
Yes	12	0.41	0.14 to 0.68			
No	27	0.34	0.17 to 0.51			
Stand alone vs. number cruncher				.70	0.15	−0.63 to 0.92
Stand alone	12	0.43	0.11 to 0.74			
Number cruncher	14	0.34	0.06 to 0.63			
Publication type: Dissertation				.75	0.13	−0.64 to 0.89
Yes	8	0.38	0.06 to 0.70			
No	37	0.32	0.18 to 0.46			
General exam assessment				.81	0.09	−0.64 to 0.82
Yes	9	0.37	0.06 to 0.67			
No	36	0.33	0.17 to 0.48			
Web communication tools[a]				.82	0.07	−0.59 to 0.74
Yes	13	0.34	0.10 to 0.58			
No	26	0.37	0.19 to 0.55			
Number cruncher tools[a]				.85	0.06	−0.59 to 0.72
Yes	14	0.34	0.10 to 0.58			
No	25	0.37	0.19 to 0.55			
Prior statistical background				.87	0.05	−0.58 to 0.67
Yes	23	0.34	0.13 to 0.59			
No	17	0.31	0.09 to 0.54			
Unit of assignment[a]				.87	0.05	−0.54 to 0.63
Individual	22	0.32	0.12 to 0.41			
Classroom or group	23	0.34	0.17 to 0.51			
Random assignment[a]				.97	0.01	−0.61 to 0.63
Yes	16	0.33	0.09 to 0.57			
No	27	0.34	0.17 to 0.51			
Number cruncher vs. Web communication[a]				.99	0.00	−0.75 to 0.76
Number cruncher	14	0.34	0.09 to 0.59			
Web communication tools	13	0.34	0.09 to 0.59			

Note. Total *k* does not always equal 45 because of missing data. The *p* value is computed from the *Q* statistic, which gives a two-tailed test when there are two groups.

[a]Study descriptor identified as statistically significant ($p < .05$) under a fixed effects model with un-Winsorized effect size weights.

[b]Study descriptor identified as statistically significant ($p < .05$) under a fixed effects model with a single Winsorized effect size weight (i.e., 1399.25 changed to 97.76).

received computer-based instruction substantially outperformed students who received lecture-based instruction when the former received more overall instruction.

Academic level. The confidence interval of each academic level subgroup ($CI._{95} = 0.41 \leq \mu \leq 0.94$ for graduates; $CI._{95} = 0.12 \leq \mu \leq 0.39$ for undergraduates) indicates that, regardless of academic level, computer-assisted instruction yields larger effects than does non-computer-assisted instruction. Moreover, studies that used graduate students yielded a larger mean effect ($d = 0.68$) than studies with undergraduate students ($d = 0.25$); this difference was found to be statistically significant, $Q(1) = 7.72$, $p = .006$. The $d_{diff} = 1.08$ indicates that an estimated 84% of studies using graduate students yielded larger effects than the average study using undergraduate students. It is apparent that although both academic levels benefit from computer-assisted instruction, graduate students seem to benefit more from such resources than do undergraduate students.

Embedded assessment. We found an advantage for computer-assisted instruction, regardless of whether studies employed an embedded assessment ($CI._{95} = 0.36 \leq \mu \leq 0.99$ for embedded assessment; $CI._{95} = 0.12 \leq \mu \leq 0.40$ for nonembedded assessment). In addition, we found a statistically significant difference between these two methods, $Q(1) = 5.36$, $p = .02$, indicating that studies making use of embedded assessment yield larger effects than those not utilizing embedded assessment. The corresponding d_{diff} (0.92) is clearly robust; an estimated 82% of studies employing embedded assessments obtain larger effects than the average study not employing embedded assessment. Thus, although both embedded and nonembedded assessments demonstrate a decided gain over lecture-only instruction, instruction that includes embedded assessment is associated with greater learning than studies not making use of such an assessment approach.

Publication type. We compared three types of publications: journal articles, papers stemming from conference presentations, and dissertations. Our findings indicate that both journal articles and dissertations report a performance advantage for computer-assisted instruction over face-to-face instruction ($CI._{95} = 0.22 \leq \mu \leq 0.53$ for journals; $CI._{95} = 0.06 \leq \mu \leq 0.70$ for dissertations). However, this does not appear to be the case for conference papers ($CI._{95} = -0.31 \leq \mu \leq 0.40$). This difference becomes even more apparent when comparing the average effect for conference papers to the collective average effect of journal articles and dissertations, $Q(1) = 2.69$, $p = .09$; $d_{diff} = .71$.

Implementation of tool. The confidence interval of each subgroup ($CI._{95} = -0.08 \leq \mu \leq 0.42$ for replacement; $CI._{95} = 0.24 \leq \mu \leq 0.54$ for supplement) indicates that the advantage of computer-assisted instruction over face-to-face instruction tended to be larger when the computer-assisted tool was used to supplement, rather than replace, face-to-face instruction. The 13 studies utilizing tools as supplements yielded a larger mean effect ($d = 0.39$) than the 32 studies utilizing such tools as replacements ($d = 0.17$), although this difference was not statistically significant, $Q(1) = 2.07$, $p = .15$; $d_{diff} = .47$. This means that an estimated 68% of studies employing the computer-aided tools as supplements yielded larger effects than the average

study utilizing those tools as replacements. An effect size this large is meaningful, suggesting that computer-assisted instruction is likely to be more effective when it is used to supplement face-to-face instruction rather than replace it.

Remaining categorical variables. None of the 19 remaining categorical variables attained statistical significance as a moderator of the relationship between computer-based instruction and student performance. However, the average effect size *d* was positive for every subgroup, and both limits of the 95% CI for *d* were greater than zero for virtually all of the subgroups (see Table 2). These data indicate computer-assisted instruction showed a positive effect across a wide range of conditions. The d_{diff} statistic exceeded 0.15 for 11 of these 19 variables, pointing toward several potentially important moderator variables.

Moderator Analysis With Continuous Variables

As described earlier, the authors rated the extent to which each treatment condition embodied each of 11 attributes such as student engagement, student control over the learning process, and the nature of feedback. As shown in Table 3, the correlations between these ratings and the observed treatment effect sizes were all positive, ranging from .02 to .20, indicating that all of the relationships were in the expected direction in this set of studies. Surprisingly, however, no individual attribute attained statistical significance. The largest correlations suggested greater advantages of computer-assisted instruction for more complex statistical concepts, with more rapid feedback, for programs that stimulated more active learning, and for programs that used more simulation (see Table 3).

Discussion

Conclusions Regarding Computer-Assisted Statistics Instruction

The average effect size of 0.33 for the 45 studies in this review demonstrates that computer-assisted instruction is, on average, more effective than lecture-only instruction. Assuming normal distributions within groups, an effect size of 0.33 indicates that 63% of students receiving computer-assisted instruction in statistics demonstrate greater learning than the average student receiving only lecture-based instruction. Equivalently, an average student (i.e., 50th percentile) who receives computer-assisted instruction in statistics demonstrates greater achievement than 63% of students who receive only a lecture version of the same material. Moreover, the 95% confidence interval for the average effect size ranges from 0.20 to 0.46. This confidence interval includes the average effect sizes reported by Hsu (2003) and Schenker (2007), 0.43 and 0.24, respectively. Thus, as in previous meta-analytic work, these findings demonstrate that computer-assisted instruction has a modest yet meaningful, beneficial impact on statistics learning.

The search for characteristics of computer-based instruction that are associated with learning outcomes revealed several potential moderators. One such study characteristic was the amount of time spent with the learning material. Specifically, computer-based treatment groups showed a greater performance advantage over lecture-only groups when the former received more overall instruction than the latter. The d_{diff} value of 1.99 (95% CI ranging from 1.03 to 2.94) indicates that the effect of additional time is large and potentially very large. Although this finding

TABLE 3

Mixed effects moderator analysis examining the correlations between weighted effect sizes and learner-centered characteristics

Learner-centered characteristic	*k*	*r*	*p*
Interactivity			
Active learning	41	.16	.28
Thoughtful responses	43	.10	.49
Learner control	43	.06	.72
Nature of feedback			
Immediacy of feedback	37	.18	.27
Targeted feedback	38	.08	.60
Sources of instructional content			
Student–student interaction	41	.13	.39
Instructor involvement	44	.12	.42
Technology's contribution beyond instructor or students	44	.05	.75
Miscellaneous			
Complexity of statistical concepts	45	.20	.16
Degree of simulation	44	.16	.27
Breadth or range of statistical topics	45	.02	.91

Note. *k* does not always equal 45 because of missing data.

is not surprising, from a practical standpoint it indicates that extra time spent with computer-assisted instruction can be highly beneficial.

The effect of computer-based instruction was significantly larger for graduate students than for undergraduates; the d_{diff} value of 1.08 (95% CI = 0.32 to 1.83) also indicates that this effect probably is large. This finding suggests that prior knowledge or expertise may be conducive to mastering statistical concepts via computer-based statistics learning tools. Previous work has found graduate students to possess more favorable attitudes toward computer-assisted instruction than undergraduates (Koroghlanian & Brinkerhoff, 2007; Yilmazel-Sahin, 2009); it is possible that such perceptions enhance the learning benefits of computer-assisted instruction. It is also possible that graduate students may be better self-regulated learners than undergraduate students. Effective self-regulated learning is a cyclical process in which the learners must manage, plan, and control their learning by setting goals and enacting strategies to achieve those goals (Borkowski & Burke, 1996; Moos & Azevedo, 2008). Several studies have shown that low knowledge learners, who tend not to self-regulate their learning effectively, may require more support or scaffolding (Azevedo, 2005; Chen, Fan, & Macredie, 2006; Moos & Azevedo, 2008). In contrast, learners with high prior domain knowledge have demonstrated more planning and monitoring behaviors during learning (Moos & Azevedo, 2008). Given their greater educational experience compared to undergraduates, graduate students may possess a greater ability to implement effective learning strategies and make better use of feedback when working computer-based tools.

We also found effect size differences related to the type of outcome measure employed by studies. Specifically, we found that embedded assessments

composed of target items contained within a broader assessment (e.g., final exam) yielded significantly larger effects than studies not utilizing this assessment approach. One possible explanation is that larger effects may be found for items more highly linked to the computer-based tool in question. However, there was no advantage for studies that used assessment specifically designed to measure the learning that was targeted by the computer-based tool (i.e., dedicated assessment). Another possibility is that, in the context of embedded assessment, the nontarget items enhanced performance on the target items. That is, working through nontarget items may prime students to think about the knowledge measured by the target items; nontarget items may activate knowledge that is related to the specific knowledge under study. For instance, if target items are designed to measure students' understanding of hypothesis testing, exposure to other items on the same exam measuring understanding of related concepts such as the sampling distribution of the mean may activate knowledge that is relevant to performing well on measures of hypothesis testing. Recent work has shown how testing can promote more effective learning and teaching (Roediger & Karpicke, 2006), but more research is needed to determine the best ways to use assessment in the context of computer-assisted instruction.

Several additional study characteristics that did not attain statistical significance for the test of differences between effect sizes nevertheless showed d_{diff} values and confidence intervals that suggest potential importance. Using computer-based tools to supplement, rather than replace, face-to-face instruction tended to yield larger performance gains (d_{diff} = 0.47, 95% CI = –0.18 to 1.13). These findings suggest that students learn best when they are exposed to an instructional approach that incorporates elements of both face-to-face and computer-based instruction.

We did not, however, obtain particularly robust findings with regard to the three technology types (i.e., stand-alone, number-cruncher, and web-based communication tools); our findings indicate that all technology types are comparably beneficial to learning and that all three hold a performance advantage over lecture-based instruction. Although we expected that the added interactivity and engagement offered by stand-alone tools would translate into larger effects when compared to number crunchers and web-based communication tools, our findings do not support this view. It is possible that stand-alone tools do not offer a learning advantage over number crunchers or web-based communication tools because they do not offer learners the greater degree of interactivity or feedback that is often ascribed to them. To examine the issue further, we conducted secondary data analyses to examine the possibility that the three types of technology differed on specific learner-centered features such as interactivity and feedback. We found that although number crunchers and web-based communication tools did not differ along any of the learner-centered attributes that we examined, stand-alone tools were found to possess significantly higher levels of interactivity (i.e., active learning, thoughtful responses, learner control) and feedback (i.e., immediacy of feedback, targeted feedback). All p values were less than .05, and effect size d values ranged from 1.04 to 3.17. In addition, relative to both number crunchers and web-based communication tools, stand-alone tools were significantly more likely to be associated with lower levels of instructor involvement (d = 1.67) and offered a significantly greater contribution to learning beyond the instructor and students (d = 2.36).

These robust findings indicate that stand-alone tools possess the key characteristics most commonly attributed to them—they offer students greater cognitive engagement and contribute positively to the overall instructional experience students have in courses employing such tools. However, despite this apparent advantage over number crunchers and web-based communication tools, our findings do not lend support to the notion that these differences translate into greater learning. One possibility is that despite offering students greater interactivity and feedback, stand-alone tools' apparent benefits can be fully harnessed only by proficient learners or experts and not by the novice learner (Larreamendy-Joerns & Leinhardt, 2006; Moos & Azevedo, 2008). That is, the various features that stand-alone tools offer may be apparent only to the experienced student and may simply be too complex or unclear for the beginning student to take full advantage. Thus, although stand-alone tools may serve as the "gold standard" of computer-based resources, one must be mindful of the importance of making such tools as clear and accessible as possible for novice learners. In light of our finding that tools used to supplement rather than replace face-to-face instruction are associated with larger effects, coupled with our finding that all technology types are equally beneficial to learning, we are left to conclude that all three types of tools can have a meaningful impact on learning when they are used to supplement, rather than substitute, face-to-face instruction. Clearly, this is an area of research that merits further examination.

As discussed earlier, previous work suggests that learner-centered characteristics (i.e., interactivity and feedback) substantially account for the benefits that students derive from computer-assisted instruction. In contrast, our findings do not support the view that the level of implementation of these features has a discernable impact on student learning. It is possible that instructional facets, such as learner control, interactivity, and feedback, do not underlie (at least to a meaningful degree) the benefits of computer-assisted instruction. Before coming to any definitive conclusions, there are two points to consider. First, it should not be overlooked that all of our findings are in the direction predicted from previous research—although failing to achieve statistical significance, such consistency points to potentially meaningful effects that are at least partially obscured by our conservative analytical approach. We were able to reliably identify the differences in the studies on the basis of these characteristics; however, the fact that these studies varied along many other characteristics clearly warranted the use of a mixed effects procedure, a decidedly conservative approach.

Another point to consider is that none of the studies included in the present meta-analysis was designed to test specific learner-centered variables; as a result, none offered tightly controlled comparisons to test those variables. Other meta-analyses that examined feedback also have produced inconclusive findings, suggesting that the effect of feedback may be obscured by variability among other factors. For instance, J. Kulik and Kulik (1988) found that immediate feedback was more effective than delayed feedback in general. However, they were surprised that there were not more studies relating computer-based instruction and feedback timing. They also acknowledged the need for more controlled studies exploring variables that may affect the effectiveness of feedback. In this way, our findings point to potentially meaningful directions for future research. For instance, although our findings do not identify feedback and interactivity as significant moderators, it is possible that the effectiveness of feedback is dependent on the

learners' self-regulatory strategies (especially with regard to self-monitoring, goal setting, and awareness of how to obtain more information) in using the feedback to their advantage (Hattie & Timperley, 2007). Although greater learner control may facilitate learning for the highly motivated and skilled students, it may hamper learning among those less motivated or those less proficient (Chen et al., 2006; Milheim & Martin, 1991). One reason we may not have obtained robust findings in examining various instructional features is that their effectiveness is contingent on learner characteristics, including self-regulation of learning skills and motivation. Indeed, prior knowledge may also affect the effectiveness of computer-based tools, as suggested by our findings that graduate students gained more from using computer-based tools than did nongraduate students.

Previous research has shown that cognitive engagement promoted by interactivity of instruction has a positive effect on the effectiveness of computer-based instruction. Bernard et al.'s (2009) meta-analysis showed that both high and moderate levels of student–content interaction were more positively associated with achievement than low levels; no such relationships between effect size and student–teacher or student–student interactions were found. These findings suggest that distance education courses that promote cognitive engagement and active learning through design of the content may be particularly effective. Although the correlations between learning and interactivity ratings were positive in our meta-analysis (see Table 3), our findings failed to replicate Bernard et al.'s significant findings regarding interactivity. It is unclear whether they used a fixed effects or mixed effects model in conducting their moderator analyses, and thus it is difficult to determine whether their tests were less conservative than ours. However, their significant finding of student–content interactivity (and indirectly active learning) may be because of their approach, which may be more sensitive in two respects. Instead of merely rating the experimental condition, they rated how the experimental and treatment conditions differed on each interactivity dimension. Moreover, they compared only distance education conditions that differed on some specific instructional aspect or aspects rather than comparing distance education with traditional instruction conditions, which may vary along multiple dimensions. Thus, they may have avoided or minimized the high level of heterogeneity observed in the current meta-analysis. Bernard et al.'s meta-analysis underscores the importance of using more sensitive approaches to test the impact of learner-based features in computer-assisted instruction by designing studies that focus on the specific features of interest. Focused research could provide tests of interactions of design features with learner characteristics.

Connection to Previous Findings and Methodological Considerations

The U.S. Department of Education, Office of Planning, Evaluation, and Policy Development (2009) meta-analysis comparing face-to-face instruction and online learning corroborates several of our key findings. Their average *d* of 0.24 falls within our 95% confidence interval for the average *d*, from 0.20 to 0.46. Their meta-analysis also parallels our findings concerning time spent on learning and the use of technology as a supplement to traditional instruction. Instruction that combined both online and face-to-face instruction had an average *d* of 0.35, larger than the average *d* for purely online instruction, 0.14. When comparing online learners who spent more time on instruction than those in the face-to-face condition, the

average *d* was 0.46, greater than the average *d* of 0.19 observed when comparing online learners who spent relatively less time or equal time to learners using traditional instruction.

In our meta-analysis, studies of higher methodological quality tend to show weaker effects than studies of lower quality. Studies that included a pretest averaged smaller effect sizes than those with no pretest (d_{diff} = 0.44, 95% CI = –0.21 to 1.09), and studies with statistical control of preexisting differences averaged smaller effects than those studies that did not use such control (d_{diff} = 0.28, 95% CI = –0.34 to .90). Papers presented at conferences tended to show weaker effects of computer-based tools than publications or dissertations (d_{diff} = 0.71, 95% CI = –0.24 to 1.65), suggesting a publication bias in favor of studies reporting statistically significant findings. Although these differences did not attain statistical significance, the consistent pattern across variables and the moderate to large values of d_{diff} suggest that meaningful effects are more likely than null effects. Mixed effects meta-analysis with the current data is not sensitive enough to provide adequately precise estimates of these effect sizes. The U.S. Department of Education, Office of Planning, Evaluation, and Policy Development (2009) meta-analysis, which employed a mixed effects model for the moderator analysis, also failed to find many statistically significant methodological moderators. They examined the following six methodological moderators: (a) sample size, (b) knowledge type tested, (c) quality of study design, (d) unit of assignment to condition, (e) instructor equivalence across conditions, and (f) equivalence of curriculum and instructional approach across conditions. The only significant moderator was whether the curriculum and instruction were equivalent between conditions. Effect sizes were smaller when the curriculum and instruction were identical or almost identical in online and face-to-face conditions (average *d* = 0.20) than when various instructional features differed between the two conditions (average *d* = 0.42). As was the case in reviews by Hsu (2003) and Schenker (2007), we also were unable to uncover many statistically robust study characteristics associated with achievement outcomes. Our analyses led us to conclude that the studies that were combined to produce these statistics represent not one but many populations. Indeed, the individual studies in this literature vary in so many ways that when they are grouped by any one index that we considered, the studies within each group are still highly heterogeneous. These studies spanned over three decades and employed wide-ranging technologies applied in diverse settings. Furthermore, these studies employed a large array of outcome measures: Some used final course grades, others used final exams scores, and still others used brief but focused quizzes. Despite our efforts at defining relatively homogenous subsets of studies, it is clear that there still remains a high degree of heterogeneity within groups of studies.

An analysis of heterogeneity showed that the observed variability among the effect sizes is larger than one would expect because of sampling error of students alone. Although the significance testing approach to evaluating heterogeneity allows one to detect whether heterogeneity is present, it does not provide an index of the amount of existing heterogeneity. To address this limitation, Hedges and Pigott (2001) suggested a simple measure of heterogeneity: the ratio of the between-variance component, τ^2, to the observed variance of the weighted mean effect size. They offered rules of thumb for interpreting this ratio; specifically, they suggested that a ratio of .33 indicates a small degree of heterogeneity, a ratio

around .67 indicates a moderate degree of heterogeneity, and a ratio around 1.0 indicates a large degree of heterogeneity. For the current meta-analysis, this ratio is .12 / .19 = .64, reflecting a moderate amount of heterogeneity among the effect sizes. This finding further corroborates our conclusion regarding the amount of heterogeneity that exists in this area of research.

The observed level of heterogeneity, coupled with the use of a mixed effects model, contributes to low statistical power for detecting the effects of moderator variables (Hedges & Pigott, 2004). Currently, the vast majority of meta-analyses employ a fixed effects approach (Schmidt, 2010; Schmidt, Oh, & Hayes, 2009). In fact, of 199 meta-analyses published in *Psychological Bulletin* as of January 2006, only 13 employed a random effects approach (Schmidt et al., 2009). There is a recognized bias against conservative methods because studies with statistically significant findings are more likely to be published than studies that fail to find statistically significant effects (Lipsey & Wilson, 1993). Given that mixed effects procedures result in reduced statistical power to identify moderators (the primary focus in most meta-analyses), researchers may be leery of employing such techniques for fear that they would undercut the likelihood of publication for their findings. This dilemma further strengthens the need to utilize alternative statistical indices that are not driven by sample size, such as d_{diff}. Indeed, our findings point to d_{diff} as a particularly meaningful index that demonstrates the strength of the relationship between a given study characteristic (i.e., potential moderator) and the study effects. Unlike the hypothesis testing approach that is so often relied on when conducting moderator analyses, d_{diff} is not influenced by the number of identified studies. A mixed effects approach offers an appropriate model that allows generalization of findings to a larger, more representative body of research and, coupled with the use of d_{diff}, is conducive to the identification of representative ranges for the effects under study.

Findings Based on a Fixed Versus Mixed Effects Model

Although a mixed effects model is clearly warranted in this area of research, the reader may wonder how our findings would differ had we opted in favor of the traditionally employed fixed effects procedure. To examine the difference between these approaches, we also conducted fixed effects moderator analyses of the categorical study descriptors shown in Table 2. With fixed effects analyses, 19 of 24 study descriptors were identified as statistically significant ($p < .05$). However, the effect size weight for one study (1399.25; reported in Hilton & Christensen, 2002) was identified as an outlier in a distribution with a mean equal to 50.94 and median equal to 15.76. As mentioned earlier, findings stemming from a fixed effects moderator analysis—in contrast to a mixed effects analysis—are greatly influenced by the magnitude of effect size weights, which are based on study sample sizes (Borenstein et al., 2009; Lipsey & Wilson, 2001). Given the potential bias resulting from the inclusion of an extreme outlier in fixed effects analyses, the extreme weight was set equal to the second largest observed weight (97.79) and fixed effects moderator analyses were conducted a second time. As noted in Table 2, fixed effects analyses with Winsorized weights produced statistical significance in 8 of the 24 categorical descriptors. The pattern of findings with Winsorized weights are consistent with our mixed effects findings in that these fixed effects findings point to statistical significance ($p < .05$) among the seven study characteristics with

the smallest p values in our mixed effects analyses. These are also the seven study characteristics with the largest values of d_{diff}, ranging from 0.33 to 1.99. To ensure that the outlier did not have an undue influence on our mixed effects findings, we also reran those analyses with the Winsorized weight. As expected, our mixed effects findings with respect to reported effect sizes (d), p values, and confidence intervals were essentially unchanged, with only a few small differences at the hundredths decimal place.

The fixed effects findings further highlight the differences in statistical power associated with the use of fixed versus mixed effects procedures. Although all of the fixed effects findings pointed to statistical significance among the categorical variables, they also showed statistical significance of the residual pooled within-group variances. The former indicates that differences exist among the subgroups under comparison, establishing them as moderator variables; the latter indicates that these moderators do not adequately account for the observed heterogeneity among the effect sizes, confirming the need for a mixed effects approach that models such heterogeneity (Lipsey & Wilson, 2001). It should be noted that although our findings are generally consistent with those obtained by Hsu (2003) and Schenker (2007), there are important methodological differences between the meta-analyses. First, Hsu employed a fixed effects approach in estimating the mean effect size and in conducting the moderator analyses; on the other hand, Schenker employed a random effects model in estimating population parameters and a mixed effects approach in conducting all moderator analyses. However, Schenker also based his findings on 117 effects stemming from 46 studies, compromising the statistical independence that should exist between the effect sizes under study (Lipsey & Wilson, 2001). Although procedures exist for dealing with dependent effect sizes (Cheung & Chan, 2004; Gleser & Olkin, 1994), none was employed. Given the pivotal role played by heterogeneity among studies examining the effectiveness of computer-assisted instruction in statistics, future analyses in this area should include an index of heterogeneity, such as the one discussed earlier (Hedges & Pigott, 2001).

Conclusions

There is compelling evidence that students using computer-based tools in their statistics courses generally demonstrate greater achievement than their peers who receive only face-to-face instruction. This conclusion applies to a wide range of implementations that vary based on the type of computer-based tool, the concepts that are addressed, the measures that are used, and the number of computer-assisted sessions. The current meta-analysis provides suggestive evidence regarding various potential moderating study characteristics. The observed level of heterogeneity across studies highlights the need for more focused studies and greater collaboration and consistency in methodological and theoretical approaches to the study of features in computer-assisted statistics instruction.

Collectively, work comparing computer-assisted instruction to traditional face-to-face instruction has demonstrated a performance advantage in favor of technology-based learning. Although the current study highlights the need for further examination of the factors that account for the achievement differences observed between the two instructional approaches, we believe that the field is currently undergoing a shift toward research that directly compares different kinds of

technological tools. Using technology in the statistics classroom is no longer novel; certainly computers are used frequently for computations and presentations as well as for communication. Increasingly, instructors are not interested in knowing merely whether computer-assisted instruction benefits students; rather, they seek to ascertain which instructional tools foster the most learning. This empirical shift is apparent in Bernard et al.'s (2009) meta-analysis and is seen among recent primary studies specific to statistics education that compare the effectiveness of new computer-based instructional tools (viz., stand-alone tools) to more typical computer-based tools, such as PowerPoint slides or commercial statistical software packages (Christou, Dinov, & Sanchez, 2007; Dinov & Sanchez, 2006; Liu, Lin, & Kinshuk, 2010). Such work reflects the understanding that computer-assisted instruction is now a staple in our statistics education. Given the increased emphasis on stand-alone tools and distance learning, research on the role of interactivity, engagement, and feedback takes on increased importance as educators continue work on improving the efficacy of technology-based statistics instruction. The high level of heterogeneity among the studies reviewed here suggests that research on computer-assisted instruction in statistics would benefit from closer collaboration on measurement and design across researchers and more research focused closely on specific features of interest.

References

**References marked with an asterisk indicate studies included in the meta-analysis.*

*Aberson, C. L., Berger, D. E., Healy, M. R., Kyle, D. J., & Romero, V. J. (2000). Evaluation of an interactive tutorial for teaching the central limit theorem. *Teaching of Psychology*, *27*, 289–291.

*Aberson, C. L., Berger, D. E., Healy, M. R., & Romero, V. J. (2002). An interactive tutorial for teaching statistical power. *Journal of Statistics Education*, *10*. Retrieved from http://www.amstat.org/publications/jse/v10n3/aberson.html

*Aberson, C. L., Berger, D. E., Healy, M. R., & Romero, V. J. (2003). Evaluation of an interactive tutorial for teaching hypothesis testing concepts. *Teaching of Psychology*, *30*, 75–78.

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

*Athey, S. (1987). *A mentor system incorporating expertise to guide and teach statistical decision making* (Doctoral dissertation). (ProQuest Digital Dissertations Retrieved from AAT 8711623)

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist*, *40*, 199–209.

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. T. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*, 213–238.

*Basturk, R. (2005). The effectiveness of computer-assisted instruction in teaching introductory statistics. *Educational Technology & Society*, *8*, 170–178.

*Benedict, J. O., & Anderson, J. B. (2004). Applying the just-in-time teaching approach to teaching statistics. *Teaching of Psychology*, *31*, 197–199.

Bernard, M., Abrami, P. C., Borokhovski, E., Wade, C. A., Tamim, R. M., Surkes, M. A., & Bethel, E. C. (2009). A meta-analysis of three types of interaction treatments in distance education. *Review of Educational Research*, *79*, 1243–1289.

*Bliwise, N. G. (2005). Web-based tutorials for teaching introductory statistics. *Journal of Educational Computing Research*, *33*, 309–325.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Suffix, UK: Wiley.

Borkowski, J. G., & Burke, J. E. (1996). Theories, models, and measurements of executive functioning: An information processing perspective. In G. R. Lyon, & N. A. Krasnegor (Eds.), *Attention, memory, and executive function* (pp. 235–261). Baltimore, MD: Brookes.

Cafolla, R. (2006). Project MERLOT: Bringing peer review to web-based educational resources. *Journal of Technology and Teacher Education*, *14*, 313–323.

Chen, S. Y., Fan, J. P., & Macredie, R. D. (2006). Navigation in hypermedia learning systems: Experts vs. novices. *Computers in Human Behavior*, *22*, 251–266.

Cheung, S. F., & Chan, D. K. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology*, *89*, 780–791.

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*, 73–105.

*Christmann, E. P., & Badgett, J. L. (1997). The effects of computers and calculators on male and female statistics achievement. *Journal of Computing in Higher Education*, *9*, 49–58.

Christou, N., Dinov, I. D., & Sanchez, J. (2007, August). *Design and evaluation of SOCR tools for simulation in undergraduate probability and statistics courses*. Paper presented at the 56th session of the International Statistical Institute, Lisbon, Portugal.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York, NY: Academic Press.

*Dimitrova, G., Maisel, R., & Persell, C. H. (1993). Using and evaluating ISEE, a new computer program for teaching sampling and statistical inference. *Teaching Sociology*, *21*, 341–351.

Dinov, I., & Sanchez, J. (2006, July). *Assessment of the pedagogical utilization of the statistics online computational resource in introductory probability courses: A quasi-experiment*. Paper presented at the 7th International Conference on Teaching Statistics, Salvador, Bahia, Brazil.

*Dixon, P. N., & Judd, W. A. (1977). A comparison of computer-managed instruction and lecture mode for teaching basic statistics. *Journal of Computer-Based Instruction*, *4*, 22–25.

*Dorn, M. J. (1993). *The effect of an interactive, problem-based hypercard modular instruction on statistical reasoning* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations (AAT 9505353)

*Dutton, J., & Dutton, M. (2005). Characteristics and performance in an online section of business statistics. *Journal of Statistics Education*, *13*. Retrieved from http://www.amstat.org/publications/jse/v13n3/dutton.html

*Erwin, T. D., & Rieppi, R. (1999). Comparing multimedia and traditional approaches in undergraduate psychology classes. *Teaching of Psychology*, *26*, 58–61.

*Frederickson, N., Reed, P., & Clifford, V. (2005). Evaluating web-supported learning versus lecture-based teaching: Quantitative and qualitative perspectives. *Higher Education*, *50*, 645–664.

*Fusilier, M. R., & Kelly, E. P. (1985). An evaluation of computer-assisted instruction for learning business statistics. *AEDS Journal*, *18*, 237–242.
Garfield, J. (1994). Beyond testing and grading: Using assessment to improve students' learning. *Journal of Statistics Education*, *2*. Retrieved from http://www.amstat.org/publications/jse/v2n1/garfield.html
Garfield, J., & Chance, B. (2000), Assessment in statistics education: Issues and challenges. *Mathematics Thinking and Learning*, *2*, 99–125.
*Gilligan, W. P. (1990). *The use of a computer statistical package in teaching a unit of descriptive statistics* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations (AAT 9025197)
Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. M. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York, NY: Russell Sage.
*Gonzalez, G. M., & Birch, M. A. (2000). Evaluating the instructional efficacy of computer-mediated interactive media: Comparing three elementary statistics tutorial modules. *Journal of Educational Computing Research*, *22*, 411–436.
*Gratz, Z. S., Volpe, G. D., & Kind, B. M. (1993, March). *Attitudes and achievement in introductory psychological statistics classes versus computer supported instruction*. Paper presented at the annual conference of Teachers of Psychology, Ellenville, NY.
Halpern, D. F., & Hakel, M. D. (2003). Applying the science of learning to the university and beyond: Teaching for long-term retention and transfer. *Change*, *35*(4), 36–42.
*Harrington, D. (1999). Teaching statistics: A comparison of traditional classroom and programmed instruction/distance learning approaches. *Journal of Social Work Education*, *35*, 343–352.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112.
Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.
Hedges, L. V. (1994). Statistical considerations. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 29–38). New York, NY: Russell Sage.
Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*, 203–217.
Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*, 426–445.
*High, R. V. (1998). *Some variables in relation to students' choice of statistics classes: Traditional versus computer-supported instruction*. Retrieved from ERIC database. (ED427762)
*Hilton, S. C., & Christensen, H. B. (2002, July). *Evaluating the impact of multimedia lectures on student learning and attitudes*. Paper presented at the meeting of International Conference on Teaching Statistics, Cape Town, South Africa.
Hsu, Y. (2003). *The effectiveness of computer-assisted instruction in statistics education: A meta-analysis* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations (AAT 3089963)
*Hurlburt, R. T. (2001). "Lectlets" deliver content at a distance: Introductory statistics as a case study. *Teaching of Psychology*, *28*, 15–20.
Jolicoeur, K., & Berger, D. E. (1986). Do we really know what makes educational software effective? A call for empirical research on effectiveness. *Educational Technology*, *26*(12), 7–11.
Jolicoeur, K., & Berger, D. E. (1988a). Implementing educational software and evaluating its academic effectiveness: Part I. *Educational Technology*, *28*(9), 7–13.

Jolicoeur, K., & Berger, D. E. (1988b). Implementing educational software and evaluating its academic effectiveness: Part II. *Educational Technology*, *28*(10), 13–19.

*Jones, E. R. (1999, February-March). *A comparison of an all-web based class to a traditional class*. Paper presented at the annual conference of the Society for Information Technology & Teacher Education, San Antonio, TX.

Kennedy, C. A., Brown, N. J. S., Draney, K., & Wilson, M. (2005, April). *Using progress variables and embedded assessment to improve teaching and learning*. Paper presented at the annual meeting of the American Education Research Association, Montreal, Canada.

*Koch, C., & Gobell, J. (1999). A hypertext-based tutorial with links to the web for teaching statistics and research methods. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 7–13.

Koroghlanian, C., & Brinkerhoff, J. (2007). Online students' technology skills and attitudes toward online instruction. *Technology Systems*, *36*, 219–244.

Kulik, C. C., & Kulik, J. A. (1991). Effectiveness of computer-based instruction: An updated analysis. *Computers in Human Behavior*, *7*, 75–94.

Kulik, C. C., Kulik, J. A., & Shwalb, B. (1986). The effectiveness of computer-based adult education: A meta-analysis. *Journal of Educational Computing Research*, *2*, 236–252.

Kulik, J. A. (2002). *School mathematics and science programs benefit from instruction technology* (Info Brief NSF 03-301). Washington, DC: National Science Foundation.

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

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

*Lane, J. L., & Aleksic, M. (2002, April). *Transforming elementary statistics to enhance student learning*. Paper presented at the annual conference of the American Educational Research Association, New Orleans, LA.

Larreamendy-Joerns, J., & Leinhardt, G. (2006). Going the distance with online education. *Review of Educational Research*, *76*, 567–605.

Larreamendy-Joerns, J., Leinhardt, G., & Correador, J. (2005). Six online statistics courses: Examination and review. *American Statistician*, *59*, 240–251.

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

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Liu, T.-C., Lin, Y.-C., & Kinshuk (2010). The application of simulation-assisted learning statistics (SALS) for correcting misconceptions and improving understanding of correlation. *Journal of Computer Assisted Learning*, *26*, 143–158.

Lovett, M., & Greenhouse, J. (2000). Applying cognitive theory to statistics instruction. *American Statistician*, *54*, 196–206.

Milheim, W. D., & Martin, B. L. (1991). Theoretical bases for the use of learner control: Three different perspectives. *Journal of Computer-Based Instruction*, *18*, 99–105.

Mills, J. D. (2002). Using computer simulation methods to teach statistics: A review of the literature. *Journal of Statistics Education*, *10*. Retrieved from http://www.amstat.org/publications/jse/v10n1/mills.html

*Mills, J. D. (2004). Learning abstract statistics concepts using simulation. *Educational Research Quarterly*, *28*(4), 18–33.

Moos, D. C., & Azevedo, R. (2008). Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemporary Educational Psychology*, *33*, 270–298.
Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, *32*, 99–113.
Moreno, R., & Mayer, R. E. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology*, *97*, 117–128.
Moreno, R., & Mayer, R. E. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, *19*, 309–326.
*Morris, E. (2001). The design and evaluation of Link: A computer-based system for correlation. *British Journal of Educational Technology*, *32*, 39–53.
*Myers, K. N. (1990). *An exploratory study of the effectiveness of computer graphics and simulations in a computer-student interactive environment in illustrating random sampling and the central limit theorem* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations (AAT 9016496)
*Olsen, C. R., & Bozeman, W. C. (1988). Decision support systems: Applications in statistics and hypothesis testing. *Journal of Research in Computing Education*, *20*, 206–212.
Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, *8*, 157–159.
*Petta, N. A. (1999). *The impact of a web-based class management system on student performance and attitudes in a quantitative statistics class* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations (AAT 9951304)
*Porter, T. S., & Riley, T. M. (1996). The effectiveness of computer exercises in introductory statistics. *Journal of Economic Education*, *27*, 291–299.
*Raymondo, J. C., & Garrett, J. R. (1998). Assessing the introduction of a computer laboratory experience into a behavioral science statistics course. *Teaching Sociology*, *26*, 29–37.
Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.
Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.
Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*, 183–192.
Schenker, J. D. (2007). *The effectiveness of technology use in statistics instruction in higher education: A meta-analysis using hierarchical linear modeling* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations (AAT 328685)
Schmidt, F. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, *5*, 233–242.
Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed versus random effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*, 97–128.
Schumm, W. R., Webb, F. J., Castelo, C. S., Akagi, C. G., Jensen, E. J., Ditto, R. M., & . . . Brown, B. F. (2002). Enhancing learning in statistics classes through the use of concrete historical examples: The space shuttle Challenger, Pearl Harbor, and the RMS Titanic. *Teaching Sociology*, *30*, 361–375.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *2*, 420–428.

*Skavaril, R. V. (1974). Computer-based instruction of introductory statistics. *Journal of Computer-Based Instruction*, *1*, 32–40.

Slavin, R. E. (1990). IBM's writing to read: Is it right for reading? *Phi Delta Kappan*, *72*, 214–216.

*Smith, B. (2003). Using and evaluating resampling simulations in SPSS and Excel. *Teaching Sociology*, *31*, 276–287.

*Song, P. W. (1992). *The effects of using computer software and computer simulations to teach statistics and probability in a Korean college* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations (AAT 9237014)

*Spinelli, M. A. (2001). The use of technology in teaching business statistics. *Journal of Education for Business*, *77*, 41–44.

*Sterling, J., & Gray, M. W. (1991). The effect of simulation software on students' attitudes and understanding in introductory statistics. *Journal of Computers in Mathematics and Science Teaching*, *10*, 51–56.

*Summers, J. J., Waigandt, A., & Whittaker, T. A. (2005). A comparison of student achievement and satisfaction in an online versus a traditional face-to-face statistics class. *Innovative Higher Education*, *29*, 233–250.

Svinicki, M. D. (1999). New directions in learning and motivation. *New Directions for Teaching and Learning*, *80*, 5–28.

*Tang, H. (2003). *The interaction of cognitive style and learning environment on student performance, course satisfaction, and attitude towards computers* (Doctoral dissertation). Retrieved from ProQuest Digital Dissertations (AAT 3092453)

Timmerman, C. E., & Kruepke, K. A. (2006). Computer-assisted instruction, media richness, and college student performance. *Communication Education*, *55*, 73–104.

U.S. Department of Education, Office of Planning, Evaluation, and Policy Development. (2009). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. Washington, DC: Author.

*Utts, J., Sommer, B., Acredolo, C., Maher, M. W., & Matthews, H. R. (2003). A study comparing traditional and hybrid Internet-based instruction in introductory statistics classes. *Journal of Statistics Education*, *11*. Retrieved from http://www.amstat.org/publications/jse/v11n3/utts.html

*Wang, X. (1999). Effectiveness of statistical assignments in MPA education: An experiment. *Journal of Public Affairs Education*, *5*, 319–326.

*Ward, B. (2004). The best of both worlds: A hybrid statistics course. *Journal of Statistics Education*, *12*. Retrieved from http://www.amstat.org/publications/jse/v12n3/ward.html

*Ware, M. E., & Chastain, J. D. (1989). Computer-assisted statistical analysis: A teaching innovation? *Teaching of Psychology*, *16*, 222–227.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

*Wilmoth, J., & Wybraniec, J. (1998). Profits and pitfalls: Thoughts on using a laptop computer and presentation software to teach introductory social statistics. *Teaching Sociology*, *26*, 166–178.

Wilson, D. B. (2005). *Meta-analysis macros for SAS, SPSS, and Stata*. Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html

Yilmazel-Sahin, Y. (2009). A comparison of graduate and undergraduate teacher education students' perceptions of their instructors' use of Microsoft PowerPoint. *Technology, Pedagogy and Education*, *18*, 361–380.

Authors

GIOVANNI W. SOSA is a graduate student in social psychology at Claremont Graduate University, 150 E. 10th St., Claremont, CA 91711; e-mail: *giovanni.sosa@cgu.edu*. His research examines the link among technology, learning, attitudes, and well-being and the factors associated with student success and achievement. In addition to being a member of the Web Interface for Statistics Education (http://wise.cgu.edu) team, he is the project director of a research program at Cal State Northridge examining the impact of video game interventions on the well-being and attitudes of older adults. As a research analyst at Chaffey College, he helps faculty and staff measure student learning and attitudes and provides the campus community with empirical findings concerning a range of student-centered initiatives.

DALE E. BERGER is a professor of psychology at Claremont Graduate University, 150 E. 10th St., Claremont, CA 91711; e-mail: *dale.berger@cgu.edu*. He teaches a range of statistics, research methodology, and data analysis courses and a cognitive course in the psychology of thinking and problem solving. He has written numerous articles in the fields of research methodology, public policy, and cognitive psychology and has consulted with various organizations on issues dealing primarily with applied data analysis. He is director of the Web Interface for Statistics Education Project.

AMANDA T. SAW, a graduate student in cognitive psychology at Claremont Graduate University, 150 E. 10th St., Claremont, CA 91711; e-mail: *amanda.saw@cgu.edu*, is interested in general learning principles and applying and evaluating educational technologies. In addition to the Web Interface for Statistics Education Project, she has conducted mathematical education research at the University of California, Los Angeles Human Perception Laboratory.

JUSTIN C. MARY is a graduate student in cognitive psychology at Claremont Graduate University, 150 E. 10th St., Claremont, CA 91711; e-mail: *justin.mary@cgu.edu.* He is interested in principles of learning, critical thinking, and computer-based learning tools.