How Does Distance Education Compare to Classroom Instruction?
A Meta-Analysis of the Empirical Literature

Robert M. Bernard, Philip C. Abrami, Yiping Lou, Evgueni Borokhovski, Anne Wade, Lori Wozney, Peter Andrew Wallet, Manon Fiset and Binru Huang

Centre for the Study of Learning and Performance
Concordia University, Montreal, Canada

1Louisiana State University
Baton Rouge, LA, U.S.A.

Running Head: Meta-Analysis of the Comparative Literature of Distance Education

Paper submitted for publication to Review of Educational Research

<<PLEASE DO NOT QUOTE OR CITE WITHOUT PERMISSION>>

January 11, 2004

This study was supported by Fonds québécois de la recherche sur la société et la culture and the Social Sciences and Humanities Research Council of Canada grants to Abrami and Bernard, and Louisiana State University Council on Research Awards and Department funding to Lou. The authors express appreciation to Janette M. Barrington, Anna Peretiatkovicz, Mike Surkes, Lucie A. Ranger and Claire Feng for their assistance and contributions. Contact: Dr. Robert M. Bernard, CSLP, LB-545-5, Department of Education, Concordia University, 1455 de Maisonneuve Blvd. W., Montreal, QC, Canada H3G 1M8 or e-mail to: bernard@education.concordia.ca.
How Does Distance Education Compare to Classroom Instruction?
A Meta-Analysis of the Empirical Literature

Abstract
A meta-analysis of the distance education (DE) literature between 1985 and 2002 was conducted that addressed the following questions: a) Overall, is interactive DE as effective, in terms of student achievement, student attitudes and retention as its classroom-based counterparts? b) What is the nature and extent of the variability of the findings? c) What conditions contribute to more effective DE as compared to classroom instruction? d) How do conditions of synchronous and asynchronous DE moderate the overall results? e) To what extent do media features and pedagogical features moderate the influences of DE on student learning? f) What is the methodological state of the literature? and g) what are important implications for practice and future directions for research? In total, 232 studies containing 599 independent achievement, attitude and retention outcomes were analyzed.
Introduction

In the same way that transitions between technological epochs often breed transitional names that are shed as the new technology becomes established (e.g., the automobile was called the “horseless carriage” and the railroad train was called an “iron horse”), research in new applications of technology in education often focuses, initially, on comparisons with more established instructional applications (e.g., classroom instruction). In the 1950s and 60s, the emergence of television as a new medium of instruction initiated a flurry of research that compared it with “traditional” classroom instruction. Similarly, various forms of computer-based instruction (1970s and 80s), multi-media (1980s and 90s), teleconferencing (1990s) and distance education (spanning all of these decades) have been investigated from a comparative perspective in an attempt to judge their relative effectiveness. It is arguably the case that these comparisons are necessary for policymakers, designers, researchers and adaptors to be certain of the relative value of innovation. Questions about relative effectiveness are important both in the early stages of development and as a field matures to summarize the nature and extent of the impact on important outcomes, giving credibility to change and helping to focus it.

This study deals specifically with comparative studies of distance education. Keegan’s (1996) definition of distance education (DE) is perhaps the most commonly cited in the literature, and involves five qualities that distinguish it from other forms of instruction: a) the quasi-permanent separation of teacher and learner; b) the influence of an educational organization, both in planning, preparation and the provision of student support; c) the use of technical media; d) the provision of two-way communication; and e) the quasi-permanent absence of learning groups. This latter element has been debated in the literature (Garrison & Shale, 1987; Verduin & Clark, 1991) because it seemingly excludes many applications of DE based upon teleconferencing technologies (i.e., audio- and video-conferences) that are group-based. Some argue that DE which simply recreates the conditions of a traditional classroom misses the point because DE of this type does not support the “anytime, anywhere” objective of access to education for students who cannot be in a particular place at a particular time. However, synchronous DE does fall within the purview of current practices and therefore qualifies for consideration. To Keegan’s definition, Rekkedal and Qvist-Eriksen (2003) add the following adjustments to accommodate e-learning:

- the use of computers and computer networks to unite teacher and learners and carry the content of the course;
- the provision of two-way communication via computer networks so that the student may benefit from or even initiate dialogue (this distinguishes it from other uses of technology in education). (p. 1)

In characterizing DE, Keegan also distinguishes between “distance teaching” and “distance learning.” It is a fair distinction that applies to all organized educational events. Since learning does not always follow from teaching it is also a useful way of discussing the elements—teaching and learning—that constitute a total educational setting. While Keegan does not go on to explain, specifically, how these differ in practice, it can be assumed that characteristics or teaching designates activities that teachers engage in (e.g., lecturing, questioning, providing feedback), while learning designates activities that students engage in (e.g., taking notes, studying, reviewing, revising).
The media used in DE have undergone remarkable changes over the years. Taylor (2001) characterizes five generations of DE, largely defined with regard to the media, and thereby the range of instructional options available at the time of their prevalence. The progression that Taylor describes moves along a rough continuum of increased flexibility, interactivity, materials delivery and access, beginning in the early years of DE when it was called correspondence education (i.e., the media were print and the post office), through broadcast radio and television and on to current manifestations of interactive multi-media, the Internet, access to Web-based resources, computer-mediated communication and, most recently, campus portals providing access to the complete range of university services and facilities at a distance. Across the history of DE research, most of these media have been implicated in DE studies in which comparisons have been made to what is often referred to as “traditional classroom-based instruction” or “face-to-face.”

**Instruction, Media and DE Comparison Studies**

Clark (1983, 1994) rightly criticized early media comparison studies on a variety of grounds, the most important of which is that the medium under investigation, the instructional method that is inextricably tied to it and the content of instruction, together, form a confound that renders their relative contributions to achieving instructional goals impossible to untangle. Clark goes on to argue that it is the instructional method that is the “active ingredient,” not the medium—the medium is simply a neutral carrier of content and of method. In essence, he argues that any medium, appropriately applied, can fulfill the conditions for quality instruction, and so cost and access should form the decision criteria for media selection. Effectively, these arguments suggest that media serve a transparent purpose in DE.

Several notable rebuttals of Clark’s position have followed (Ullmer, 1994; Kozma, 1994; Morrison, 1994; Tennyson, 1994). Kozma argued that Clark’s original assessment was based on “old non-interactive technologies” that simply carried method and content, where a distinction between these elements could be clearly drawn. More recent media uses, he added, involve highly interactive sets of events that occur between learners and teachers, among learners (e.g., collaborative learning), often within a constructivist framework, and even between learners and non-human agents or tools, so that a consideration of discrete variables no longer makes sense. The distinction here seems to be “media to support teaching” and “media to support learning,” completely in line with Keegan’s reference to distance teaching and distance learning.

Cobb (1997) added an interesting wrinkle to the debate. He argued that under certain circumstances, the efficiency of a medium or symbol system can be judged by how much of the learner’s cognitive work it performs. By this logic, some media, then, do have advantages over other media, since it is “easier” to learn some things with certain media than with others. The way to advance media design, according to Cobb, “is to model learner and medium as distributed information systems, with principled, empirically determined distributions of information storage and processing over the course of learning” (p. 33). According to this argument, the medium becomes the tool of the learner’s cognitive engagement and not simply an independent and neutral means for delivering content. It is what the learner does with a medium that counts, not so much what the teacher does. It is within this distinction between teaching and learning that differences in media types used in DE might arise.

**Why Do Comparative DE Studies?**

One of the differences between DE and media comparison studies is that DE is not a medium of
depends entirely on the availability of media for delivery and communication (Keegan, 1996). DE can be non-interactive or highly interactive and may, in fact, encompass one or many media types (e.g., print + video + computer-based simulations + computer conferencing) in the service of a wide range of instructional objectives. In the same way, classroom instruction may include a wide mix of media forms. So in a well-conceived and executed comparative study, where all of these aspects are present in both conditions, differences may relate more to the proximity of learner and teacher (i.e., one of Keegan’s defining characteristics of DE) and differential means through which interaction and learner engagement can occur. Synchronicity and asynchronicity, as well as the attendant issues of instructional design, student motivation, feedback and encouragement, direct and timely communication and perceptions of isolation (Bernard & Amundsen, 1989) might then form the major distinguishing features of DE and classroom instruction. Shale (1990) comments “In sum, DE ought to be regarded as education at a distance. All of what constitutes the process of education when teacher and student are able to meet face-to-face also constitutes the process of education when teacher and student are physically separated.” (p. 334)

This, in turn, suggests that “good” DE applications and “good” classroom instruction should be, in principle, relatively equal to one another, regardless of the media used, especially if a medium is used simply for the delivery of content. However, when the medium is placed in the hands of learners to make learning more constructive or more efficient, as suggested by Kozma and Cobb, the balance of effect may shift. In fact, in DE, media may transform the learning experience in ways that are unanticipated and not regularly available in face-to-face instructional situations. For example, the use of computer-mediated communication means students must use written forms of expression to communicate with one another in articulating and developing ideas, arguing contrasting viewpoints, refining opinions, settling disputes, and so on (Abrami & Bures, 1996). This use of written language and peer interaction allows for the possibility of increased reflection (Hawkes, 2001), the development of writing skills (Winkelmann, 1995) and higher quality performance and complex problem solving skills through peer modeling and mentoring (Lou, in press; Lou, Deidic & Rosenfield, 2003; Lou & MacGregor, 2002). The critical thinking literature goes so far as to suggest that activity of this sort promotes the development of higher-order thinking skills (Garrison, Anderson & Archer, 2001; McKnight, 2001).

Is it necessary or even desirable, then, to continue to conduct studies that directly compare DE with classroom teaching? Clark (2000), by exclusion, claims that it is not: “… all evaluations should explicitly investigate the relative benefits of two different but compatible types of DE technologies found in every DE program” (p. 4). By contrast, Smith and Dillon (1999) argue that comparative studies are still useful, but only when they are done in light of a full analysis of media attributes and their hypothesized effects on learning, and when these same attributes are present and clearly articulated in the compared conditions. In the eyes of Smith and Dillon, it is only under these circumstances that comparative studies can push forward our understanding of the features of DE and classroom instruction, that make them similar and different. Unfortunately, as Smith and Dillon point out, this level of analysis and clear accounting of the similarities and differences between treatment and control is not often reported in the literature, and so it is difficult to determine whether confounds across treatments exist that would render such studies uninterpretable.

There may be a more practical reason for assessing the effectiveness of DE in comparison to its classroom alternatives. There was a time when DE was regarded simply as a reasonable alternative to campus-based education, primarily for students who had restricted access to
campuses because of geography, time constraints or other circumstances. And by virtue of the limitations of the communication facilities that existed at that time (e.g., mail, telephone, TV coverage), DE itself tended to be restricted by geographical boundaries (e.g., for many years the UK Open University was available only to students in Britain). However, the reality of “learn anywhere, anytime” promulgated largely by the communication and technological resources offered by the Internet and broadband ISPs have set traditional educational institutions into intense competition for the world-wide market of “online learners.” So it is arguable that finding answers to the question that has guided much of the comparative research on DE in the past—Is distance learning as effective as classroom learning?—has become even more pressing. Should educational institutions continue to develop and market Internet learning opportunities without knowing whether they will be as effective as their classroom-based equivalents or, in the worse case, whether they will be effective at all? Based on longstanding instructional design thinking, it is not enough to develop a technology-based course simply because the technology of delivery exists, and yet the reverse of this very thinking seems to prevail in the rush to get courses and even whole degree programs online. Beyond being simply a “proof of worthiness,” well-designed studies can suggest to administrators and policymakers not only whether DE is a worthwhile alternative, but also in which content domains, with which learners, under what pedagogical circumstances and with which mix of media the transformation of courses and programs to DE is justified. In fact, it is not unreasonable to suggest that such studies might be conducted under “local circumstances” for the primary purpose of making decisions that affect institutional growth on that particular campus.

Evidence of Effectiveness
The answer to the DE effectiveness question (or any research question for that matter) cannot be found in a single study. It is only through careful reviews of the general state of affairs in a research literature that large questions can be addressed, and that the quality of the research itself and the veracity of its findings can be assessed.

There have been many attempts to summarize the comparative research literature of DE. The most comprehensive, but least assiduous, is Russell’s (1999) collection of 355 (and counting) “no significant difference” studies. Based on evidence in the form of fragmented annotations (e.g., … no significant difference was found …) of all of the studies that could be located and contrasting this with the much smaller number of “significant difference studies” (which can be either positive or negative), Russell declared that there is no compelling evidence to refute Clark’s original claim (1983) that a delivery medium contributes little if anything to the outcomes of planned instruction, and, by extension, that there is no advantage in favor of technology-delivered DE. But there are several problems with this approach. First, not all studies are of equal quality and rigor, and to include them all, without qualification or description, renders conclusions and generalizations suspicious, at best. Second, an accepted null hypothesis does not deny the possibility that unsampled differences do exist in the population; it only means that they do not exist in the sample being studied. This is particularly true in small-sample studies where power to reject the null hypothesis (and thus the risk of making Type II errors) is high. Third, differential sample sizes of individual studies make it impossible to aggregate the results of different studies solely on the basis of their test statistics. So Russell’s work represents neither a sufficient overall test of the hypothesis of no difference nor an estimate of the magnitude of effects attributable to DE (i.e., a non-significant hypothesis test can yield a reasonable effect size in small-sample studies).
Another widely cited report (Phipps & Merisotis, 1999), prepared for the American Federation of Teachers and the National Education Association, and entitled “What’s the Difference: A Review of Contemporary Research on the Effectiveness of Distance Education in Higher Education,” appears to contain a similar level of bias to that in Russell’s work, but for a different reason. In the words of the authors, “While this review of original research does not encompass every study published since 1990, it does capture the most important and salient of these works.” (p. 154). In fact, just over 40 empirical investigations are cited to illustrate specific points that are made by the authors. The problem is, how can one judge importance or salience without carefully crafted inclusion and exclusion criteria? The bias that is risked, then, is one of selecting research, even unconsciously, to make one’s point, rather than accurately characterizing the state of the research literature around a given question. While one of the findings of the report may generally be true—that the literature lacks rigor of methodology and reporting—the findings of the “questionable effectiveness of DE” based on a select number of studies is no more credible than Russell’s claim of non-significance based on everything that has ever been published. Somewhere between these extremes lies evidence that can be taken as more representative of the true state of affairs in the population.

In addition to these, there have been a number of more or less extensive narrative reviews of research (e.g., Berge & Mrozowski, 2001; Saba, 2000; Jung & Rha, 2000; Schlosser & Anderson, 1994; Moore & Thompson, 1990). This type of research has long been known for its subjectivity, potential bias and inability to answer questions about magnitude of effects.

Meta-analysis or quantitative synthesis was developed by Gene Glass and his associates (Glass, McGaw & Smith, 1981) as an alternative to the selectivity of narrative reviews and the problem of basing conclusions on test statistics represented by different sample sizes (i.e., box score method of review). Meta-analysis makes it possible to combine studies with different sample sizes based upon a common metric, the effect size. Effect size (Cohen’s d) is a sample-size based index of standardized differences between a treatment and control group (it can also be expressed as a correlation coefficient) that can be averaged in a way that test statistics cannot. Refinements (Hedges’ g) by Hedges and Olkins (1985) further reduce the bias resulting from differential sample sizes among studies. Thus a meta-analysis is an approach to estimating how much one treatment differs from another, over a large set of similar studies and the variability that is associated with it. An additional advantage of meta-analysis is that moderator variables (i.e., coded study features) can also be investigated to explore more detailed relationships that may exist in the data. A careful analysis of the accumulated evidence allows a researcher to explore what might be responsible for the variability in findings across media, instructional design, course features, students, settings, etc., as well as research methodology, therefore shedding light on some of the issues of media, method and experimental confounds pointed out by Clark and others. At the same time, failure in one's ability to explore these issues exposes the limitations in the existing research, both in quantity and quality, indicating directions for further inquiry.

In summary, meta-analysis has the following advantages: a) answers questions about size of effect; b) explores, systematically, variability in effects size (e.g., coding of several dozen study features); c) allows for control over internal validity (i.e., comparison studies vs. one shot case studies) and, because, one looks at collection of studies, also maximizes external validity or generalizability; d) provides high in statistical power over a collection of studies; e) uses the student as the unit of analysis is the, not the study—large sample studies have higher weight; f) allows us to easily add new studies as they become available or delete anomalous studies as they are identified; f) allows us to add new study features and outcomes in future analyses and/or set.
directions for new primary research (e.g., motivation characteristics—task choice, persistence, perceived task value; pedagogical effectiveness—faculty time, student time; cost effectiveness; cost per student; levels of learning [e.g., simple knowledge or comprehension vs. higher-order thinking]); g) allows us to re-analyze or analyze parts of the data for special purposes (e.g., military studies, synchronous vs. asynchronous instruction, web-based instruction); and h) allows comment on what we know, what is new and what we need to know (Abrami, Cohen & d’Apollonia, 1988; Bernard & Naidu, 1990).

Five quantitative syntheses that are specifically related to DE and its correlates have been published (Shachar & Neumann, 2003; Ungerleider & Burns, 2003; Allen, Bourhis, Burrell & Mabry, 2002; Cavanaugh, 2001; Machtmes & Asher, 2000). In the most recent meta-analysis, Shachar and Neumann reviewed 86 studies, dated between 1990 and 2002, and found an effect size for student achievement of +0.37, which, if it holds up belies the general impression given by other studies that DE and classroom instruction are relatively equal. In another recent study, Ungerleider and Burns did a systematic review for the Council of Minister of Education, Canada (CMEC), including a quantitative meta-analysis of the literature of networked and online learning (i.e., not specifically DE). They found poor methodological quality, to the extent that only 12 achievement and four satisfaction outcomes were analyzed. They also found an overall effect size of zero for achievement and an effect size of –0.509 for satisfaction. Both findings were significantly heterogeneous. Here is an example of two credible works providing conflicting evidence as to the state of comparative studies.

Allen, Bourhis, Burrell and Mabry summarized 25 empirical studies in which DE and classroom conditions were compared on the basis of measures of student satisfaction. Studies were excluded from consideration if they did not contain a comparison group and did not report sufficient statistical information from which effect sizes could be calculated. The results revealed a slight correlation ($r = +0.031$, $k = 25$, $N = 4,702$, significantly heterogeneous sample) favoring classroom instruction. When three outliers were removed from the analysis, the correlation coefficient increased to 0.090, and the homogeneity assumption was satisfied. Virtually no effects were found for “channel of communication” (video, audio and written) or its interaction with “availability of interaction.” This meta-analysis is limited in several ways: a) it investigates only one outcome measure, student satisfaction, arguably one of the least important indicators of effectiveness; and b) its sample size and range of coded moderator variables yield little more than basic information related to the question of DE effectiveness.

The Cavanaugh meta-analysis examined interactive (i.e., videoconferencing and telecommunications) DE technologies in K–12 learning in 19 experimental and quasi-experimental studies on the basis of student achievement. Studies were selected on the following bases: a) they included a focus on interactive DE technology; b) they were published between 1980 and 1998; c) they included quantitative outcomes from which effect sizes could be extracted; and d) they were free from obvious methodological flaws. In 19 studies ($N = 929$) that met these criteria, results indicated an overall effect size (i.e., weighted mean difference) of +0.015 in favor of DE conditions for a significantly heterogeneous sample. This effect size was considered to be not significant. Subsequent investigation of moderator variables revealed no additional findings of consequence. This study is limited in its purview to K-12 courses, generalizing to what is perhaps the least developed “market” for DE.

The fourth meta-analysis performed by Machtmes and Asher compared live or pre-produced adult tele-courses with their classroom equivalents on measures of classroom achievement in either experimental or quasi-experimental designs. Out of 30 studies identified, 19 studies dated
between 1943 and 1997 were coded for effect size and study features. The overall weighted effect size for these comparisons was –0.0093 (not significant; ranging from +1.50 to –0.005). The assumption of homogeneity of effect size was violated, and was attributed to differences in learners’ level of education and differences in technology over the period of time under consideration. Three study features were found to affect student achievement: type of interaction available, type of course and type of remote site.

So in the literature of DE comparison reviews, we find only fragmented and partial attempts to address the myriad of questions that might be answerable from the primary literature, but we also find great variability among findings but general agreement concerning the poor quality of the literature. In this era of proliferation of various technology-mediated forms of DE, it is time for a comprehensive review of the empirical literature, to assess the quality of the DE research literature, systematically, and to attempt to answer questions relating to the effectiveness of DE and to suggest directions for future practice and research.

Synchronous and Asynchronous DE
In the age of the Internet and CMC there is a tendency to think of DE in terms of education, “anywhere, anytime.” DE of this type truly fits two of Keegan’s (1996) definitional criteria, “the quasi-permanent separation of teacher and learner” and “the quasi-permanent absence of learning groups.” However, much of what is called DE does not fit either of these two criteria, rendering it DE that is group-based and time and place-dependent. This form of DE, which we will call synchronous DE, is not so very different from early applications of distributed education via closed-circuit television on university campuses (e.g., Penn State) that began in the late 1940s. The primary purpose of this movement in the U.S. was to economize on teaching resources and subject matter expertise by distributing live lectures and later, mediated questioning and discussion, to many “television classrooms” or remote sites across a university campus or other satellites locales. Many studies of this form of instruction produced “no significant difference” between the live classroom and the remote site (e.g., Carpenter & Greenhill, 1955; 1958), leading to the general conclusion that:

“… there was no difference in achievement between the students in the television classes and those in the regular classes. ... the average total score for the semester also revealed no difference in achievement between the classes taught by the two methods. It is clear, however, that in many college courses students can be expected to do equally well in examinations whether they have been taught by a teacher in a regular classroom or by the same teacher over television. When the final scores were adjusted to compensate for the initial inequality, it was found that there was no difference among students ....”


The term distance education became attached to this form of instruction as the availability and reliability of videoconferencing and interactive television began to emerge in the mid-1980s. The premise, however, remains the same: two or more classes in different locations connected via some form of telecommunication technology, being directed by one or more teachers. According to Mottet (1998) and Ostendorf (1997), this form of “emulated tradition classroom instruction” is the fastest growing form of DE in U.S. universities and so it is important for us to know how it affects learners who are involved in it.

Contrasted with this “group-based” form of instruction is “individually-based” DE where students in remote locations work independently or in asynchronous groups, usually with the
support of an instructor or tutor. We have called this asynchronous because DE students are not synchronized with classroom students and because communication is largely asynchronous by e-mail or through CMC software. Chat rooms and the like, of course, offer an element of synchronicity, but this is usually an optional feature of the instructional setting. Asynchronous DE has its roots in correspondence education where learners were truly independent, connected to an instructor or tutor by the postal system; communication was truly asynchronous because of postal delay. Because of the differences in synchronous and asynchronous DE just noted, we decided to examine these two patterns undivided and divided on all of the primary measures. In fact this distinction formed a natural division around which the majority of the analyses revolved.

For some, the key definitional feature of DE is the physical separation of learners in space and time. For others the physical separation in space only is a sufficient condition for DE. In the former definition, asynchronous communication is the norm. In the later definition, synchronous communication is the norm.

We take no position on which of these definitions is correct but note that there are numerous instances in the literature where both synchronous and asynchronous forms of communication are available to the learner. We have included both types in our review in order to examine whether synchronicity affects learning. Where there exists choice in design, knowing the influence of synchronicity may guide instructional design. Where there is no choice in design and students must learn separated by both space and time, the influence of synchronicity may serve to encourage, as appropriate, the development of new instructional resources as alternative supports for student learning needs. There are, of course, hybrids of these two, referred to by some as “distributed education” (e.g., Dede, 1996). We did not attempt to separate these mixed patterns from those in which students truly worked independently from one another. So within the asynchronous studies there is an element of within-group synchronicity (i.e., DE students communicating, synchronously, among themselves). However, this does not affect the defining element of synchronicity as it is described here.

**Statement of the Problem**

The overall intention of this meta-analysis is to provide an exhaustive quantitative synthesis of the comparative research literature of DE, from 1985 to the present, across all age groups, media types, instructional methods and outcome measures. From this literature it seeks to answer the following questions:

1. Overall, is interactive DE as effective, in terms of student achievement, student attitudes and retention, as its classroom-based counterparts?
2. What is the nature and extent of the variability of the findings?
3. How do conditions of synchronous and asynchronous DE moderate the overall results?
4. What conditions contribute to more effective DE as compared to classroom instruction?
5. To what extent do media features and pedagogical features moderate the influences of DE on student learning?
6. What is the methodological state of the literature? and
7. What are important implications for practice and future directions for research?
Method

This meta-analysis is a quantitative synthesis of empirical studies since 1985 that compared the effects of DE and traditional classroom-based instruction on student achievement, attitude and student retention (i.e., opposite of drop-out). The year 1985 was chosen as a cut-off date since electronically mediated, interactive DE became widely available around that time. The procedures employed to conduct this quantitative synthesis are described below under the following subheadings: working definition of DE, inclusion/exclusion criteria, data sources and search strategies, outcomes of the searches, outcome measures and effect size extraction, study features coding and data analysis.

Working Definition of DE

The working definition of DE builds on Nipper's (1989) model of “third-generation distance learning,” as well as Keegan's (1996) synthesis of recent definitions. Linked historically to developments in technology, first generation DE refers to the early days of print-based correspondence study. Characterized by the establishment of the Open University in 1969, second generation DE refers to the period when print materials were integrated with broadcast TV and radio, audio and videocassettes, and increased student support. Third generation DE was heralded by the invention of Hypertext and the rise in the use of teleconferencing (i.e., audio and video). To this, Taylor (2001) adds the “fourth generation,” characterized by flexible learning (e.g., CMC, Internet accessible courses) and the “fifth generation” (e.g., online interactive multimedia, Internet-based access to WWW resources). Generations three, four and five represent moves away from directed and non-interactive courses to those characterized by a high degree of learner control and two-way communication, as well as group-oriented processes and greater flexibility in learning. With new communication technologies in hand and renewed interest in the convergence of DE and traditional education, this is an appropriate time to review the research on third, fourth and fifth generation DE. Our definition of DE for the inclusion of studies thus reads:

- The semi-permanent separation (place and/or time) of learner and instructor during planned learning events.
- The presence of planning and preparation of the learning materials, student support services, and the final recognition of course completion by an educational organization.
- The provision of two-way media to facilitate dialogue and interaction between the students and the instructor and among students.

Inclusion/Exclusion Criteria

To be included in this meta-analysis, each study had to meet the following inclusion/exclusion criteria:

1. Involve an empirical comparison of DE as defined in this meta-analysis (including satellite/TV/radio broadcast + telephone/e-mail, e-mail-based correspondence, text-based correspondence + telephone, web/audio/video-based two-way telecommunication) with face-to-face classroom instruction (including lectures, seminars, tutorials and laboratory sessions). Studies comparing DE with national standards or norms, rather than a control condition, were excluded.

2. Involve “distance from instructor” as a primary condition of the DE condition. DE with some face-to-face meetings (less than 50%) was included. However, studies where
3. Report measured outcomes for both experimental and control groups. Studies with insufficient data for effect size calculations (e.g., with means but no standard deviations or no inferential statistics or no sample size) were excluded.

4. Be publicly available or archived.

5. Include at least one achievement, attitude or retention outcome measure.

6. Include an identifiable level of learner. All levels of learners (from kindergarten to adults, whether informal schooling or professional training) were admissible.

7. Be published or presented no earlier than 1985 and no later than December of 2002.

8. Include outcome measures that were the same or comparable. If the study explicitly stated that different exams were used for the experimental and control groups, the study was excluded.

9. Include outcome measures that reflected individual courses rather than whole programs. Thus, programs composed of many different courses, where no opportunity existed to analyze conditions and the corresponding outcomes for individual treatments, were excluded.

10. Include only studies with an $N$ of 2 or greater (i.e., enough to form a standard deviation).

11. Include only the published source, when data about a particular study were available from different sources (e.g., journal article and dissertation). Additional data from the other source were used only to make coding study features more detailed and accurate.

Data Sources and Search Strategies

The studies used in this meta-analysis were located through a comprehensive search of publicly available literature from 1985 through December of 2002. Electronic searches were performed on the following databases: ABI/Inform, Compendex, Cambridge Scientific Abstracts, Canadian Research Index, Communication Abstracts, Digital Dissertations on ProQuest, Dissertation Abstracts, Education Abstracts, ERIC, PsycInfo, and Social SciSearch. Web searches were performed using Google, AlltheWeb, and Teoma search engines. A manual search was performed in ComAbstracts, Educational Technology Abstracts and in several distance learning journals, including The American Journal of Distance Education, Distance Education, Journal of Distance Education, Open Learning and Journal of Telemedicine and Telecare; and in several conference proceedings, including AACE, AERA, CADE, EdMedia, E-Learn, SITE, and WebNet. In addition, the reference lists of several earlier reviews including Moore and Thompson (1990); Russell (1999); Machtmes and Asher (2000); Cavanaugh (2001); Allen, Bourhis, Burrell, and Mabry (2002); and Shachar (2002) were reviewed for possible inclusions. Although search strategies varied depending on the tool used, generally search terms included: “distance education,” “distance learning,” “open learning” or “virtual university,” AND (“traditional,” “lecture,” “face-to-face” or “comparison”).

Outcomes of the Searches

In total, 2,262 research abstracts concerning DE and traditional classroom-based instruction were examined and 862 full-text potential includes retrieved. Each of the studies retrieved was read by two researchers for possible inclusion using the inclusion/exclusion criteria. The initial inter-rater
agreement as to inclusion was 89%. Any study that was considered for exclusion by one researcher was crosschecked by another researcher. Two hundred and thirty-two (232) studies met all inclusion criteria and were included in this meta-analysis; 630 were excluded. Table 1 shows the categories of exclusion and the number and percentage of the excluded studies.

Table 1. Categories of exclusions, number of studies excluded and percentages.

<table>
<thead>
<tr>
<th>Category Name</th>
<th>N Excluded</th>
<th>% Excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Review and conceptual articles</td>
<td>52</td>
<td>8.25</td>
</tr>
<tr>
<td>2. Case studies, survey results and qualitative studies</td>
<td>55</td>
<td>8.73</td>
</tr>
<tr>
<td>3. Studies with some violation of either DE or F2F definitions</td>
<td>295</td>
<td>46.83</td>
</tr>
<tr>
<td>4. Collapsed data, mixed conditions or program-based findings</td>
<td>43</td>
<td>6.83</td>
</tr>
<tr>
<td>5. Insufficient statistical data</td>
<td>97</td>
<td>15.40</td>
</tr>
<tr>
<td>6. Non-retrievable studies</td>
<td>10</td>
<td>1.58</td>
</tr>
<tr>
<td>7. “Out-of-date” studies</td>
<td>21</td>
<td>3.33</td>
</tr>
<tr>
<td>8. Duplicates</td>
<td>11</td>
<td>1.75</td>
</tr>
<tr>
<td>9. Multiple reasons</td>
<td>46</td>
<td>7.30</td>
</tr>
<tr>
<td>Total</td>
<td>630</td>
<td>100</td>
</tr>
</tbody>
</table>

Outcomes Measures and Effect Size Extraction

Outcome measures. We chose not to develop rigid operational definitions of the outcome measures, but instead used general descriptions. Achievement outcomes were objective measures—standardized tests, researcher-made or teacher-made tests or a combination of these—that assessed the extent to which students had achieved the instructional (i.e., learning) objectives of a course. While most measured the acquisition of content knowledge, tests of comprehension and application of knowledge were also included.

Attitude measures and inventories were more subjective reactions, opinions or expressions of satisfaction or evaluation of the course as a whole, the instructor, the course content or the technology used. Some attitude measures could not be classified in these terms and were labeled “other attitudes.”

Retention outcomes were measures of the number or percentage of students who remained in a course out of the total who had enrolled. When these numbers or percentages were expressed in terms of dropout, they were converted to reflect retention.

Effect size extraction. Effect sizes were extracted from numerical or statistical data contained in the study. The basic index for the effect size calculation ($d$) was the mean of the experimental group (DE) minus the mean of the classroom group divided by the pooled standard deviation (see Equation 1).

$$d = \frac{\bar{Y}_E - \bar{Y}_C}{s_{pooled}}$$

Cohen’s $d$ was converted to Hedges’ $g$ (i.e., unbiased estimate) using Equation 2 (Hedges & Olkin, p. 81).
Equation 2: \[ g \frac{1}{\sqrt{N}} (\frac{3}{4N} - 9) d \]

Effect sizes from data in forms such as \( t \)-tests, \( F \)-tests, \( p \)-levels and frequencies were computed via conversion formulas provided by Glass, McGaw and Smith (1981) and Hedges, Shymansky and Woodworth (1989). These effect sizes were referred to in coding as “estimated effect sizes.”

The following rules governed the calculation of effect sizes:

- When multiple achievement data were reported (e.g., assignments, midterm and final exams, GPAs, grade distributions), the final exam scores were used for calculating the effect size.
- When there was more than one control group and they did not differ considerably, the weighted average of the two conditions was used.
- If only one of the control groups could be considered “purely” control (i.e., classical face-to-face instructional mode), while others involved some elements of DE treatment (e.g., originating studio site), the former was used as the control group.
- In studies where there were two DE conditions and one control condition, the weighted average of the two DE conditions was used.
- In studies where instruction was simultaneously delivered to an originating site and remote sites (e.g. two-way videoconferencing), the originating site was considered to be the control condition and the remote site(s) the DE condition.
- For attitude inventories, we used the average of all items falling under one type of outcome (e.g., attitude towards subject matter) so that only one effect size was generated from each study for each outcome.
- For studies reporting only a significance level, effect sizes were estimated (e.g., \( t = 1.96 \) for \( \alpha = .05 \)).
- When the direction of the effect was not available, we used an estimated effect size of zero.
- When the direction was reported, a “midpoint” approach was taken to estimate a representative \( t \)-value (i.e., midpoint between 0 and the critical \( t \)-value for the sample size to be significant; Sedlmeier & Gigerenzer, 1989).

The unit of analysis was the independent study finding; multiple outcomes were sometimes extracted from the same study both within and between outcome types. For within outcome types (e.g., achievement), multiple outcomes were extracted for different courses; when there were several measures for the same course, the more stable outcome (e.g., posttest instead of quizzes) was extracted.

Outcomes and effect sizes from each study were extracted by two researchers, working independently, and then compared for reliability. The inter-coder agreement rate was 91% for the number of effect sizes extracted within studies and 96% for effect size calculation. In total, 688 independent effect size (i.e., 321 achievement outcomes, 262 attitude outcomes and 105 retention outcomes) were extracted.
**Study Features Coding**

*Initial coding.* A comprehensive codebook was initially developed based on several earlier narrative reviews (e.g., Phipps & Merisotis, 1999), meta-analyses (e.g., Cavanaugh, 1999), conceptual papers (e.g., Smith & Dillon, 1999), critiques (e.g., Saba, 2000) and a review of 10 sample studies. The codebook was revised as a result of sample coding and a better understanding of the literature and the issues drawn from it. The final codebook included the following categories of study features: outcome features (e.g., outcome measure source), methodology features (e.g., instructor equivalence), course design (e.g., systematic instructional design procedures used), media and delivery (e.g., use of two-way videoconferencing), demographics (e.g., subject matter), and pedagogy (e.g., problem-based learning). Of particular interest in the analysis were the related to methodology, pedagogy and media characteristics. Some study features were modified and others were dropped (e.g., type of student learning) if there was insufficient data in the primary literature for inclusion in the meta-analysis are contained the Appendix. The variables and study features that were used in the final coding. In addition to these codes, elaborate operational descriptions were developed for each item and used to guided coders.

*Operational definitions of coding options.* In order to operationalize the coding scheme and to make coding more concrete, we developed definitions of “more than,” “equal to” and “less than.” “More than” was defined as 66% or more, “equal to” as 34% to 65%, and “less than” as 33% or less. This approach to coding sets up a comparison between a DE outcome and a Control outcome within each coded item that allowed us to quantify some aspects of study features (i.e., methodology, pedagogy and media) that have heretofore been ignored or dealt with qualitatively. Thus, we hoped the meta-analysis would allow us to address the longstanding controversy regarding the effects of media and pedagogy. As well, this form of coding enabled us to estimate, empirically, the state of the DE research literature from a quality perspective. Each study was coded by two coders independently and compared. Their initial coding agreement was 90%. Disagreements between coders were resolved through discussion and further review of the disputed studies and in some difficult cases among the whole research team.

*Synchronous and asynchronous DE.* Outcomes were split for the purpose of analysis into synchronous and asynchronous DE on the basis of the study feature “SIMUL.” This study feature described whether the classroom and DE conditions met simultaneously with each other, linked by some form of telecommunication technology like videoconferencing or were separate, and therefore were not linked in any way. The term asynchronous, therefore, does not refer as much to “asynchronous communication” among instructors and/or students, as it does to the fact that it was not synchronized with a classroom. As a result of this definition, some DE students did communicate synchronously with instructors/other students, but this was not typically the case. We did not separate conditions where inter-DE synchronous communication occurred from those where it did not. Outcomes where “SIMUL” was missing were considered “unclassified” and not subjected to thorough analysis.

*Recoding methodological study features.* Thirteen coded study features relating to the methodological quality of the outcomes were recoded according to the scheme shown in Table 2. Equality between treatment and control was given a weighting of +2 and inequality was recoded as −2 to reflect this extreme discrepancy. The two indeterminate conditions (i.e., one group known and the other not known) were recoded to zero. We had three choices for dealing with the substantial amount of missing information recorded on the coding sheets: 1) use only available information and treat missing data as missing (this would have precluded multiple regression
modeling of study features, since each case had at least one study feature missing); 2) recode missing data using a mean substitution procedure under the assumption that missing data were “typical” of the average for each study feature; or 3) code missing data as zero under the assumption that it also, represents indetermination. We chose the last of these three options.

Table 2. Study feature codes and the methodological recodes assigned to them

<table>
<thead>
<tr>
<th>Study Feature Codes and their Meaning</th>
<th>Recode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DE more than control group</td>
<td>-2</td>
</tr>
<tr>
<td>2. DE reported/control group not reported</td>
<td>0</td>
</tr>
<tr>
<td>3. DE equal to control group</td>
<td>+2</td>
</tr>
<tr>
<td>4. Control reported/DE not reported</td>
<td>0</td>
</tr>
<tr>
<td>5. DE less than control group</td>
<td>-2</td>
</tr>
<tr>
<td>999. Neither DE nor Control Group used explicitly or missing</td>
<td>0</td>
</tr>
</tbody>
</table>

\(^1\)Some study had three values (1, 2 and 999) and were coded 2, 1 and 0.

The coded study features were: 1) type of publication; 2) type of measure; 3) effect size (i.e., calculated or estimated); 4) treatment duration; 5) treatment time proximity; 6) instructor equivalence; 7) selection bias; 8) time-on-task equivalence; 9) material equivalence; 10) learner ability equivalence; 11) mortality; 12) class size equivalence; and 13) gender equivalence.

Recoding pedagogical and media study features. To allow us to explore the variability among DE outcomes using multiple regression, we recoded the pedagogical and media-related study features. Using a procedure that is similar to that used to produce the methodological study features, pedagogical and media-related study features were recoded to reflect a contrast between features favoring DE conditions and features favoring classroom conditions. We faced the same problem of missing data with pedagogical and media study features as we did with methodological features. Again we chose to code missing values to zero. Our view was that this was the most conservative approach, since that gave missing values equal weight across all of the study features (i.e., mean substitution would have given unequal weight). An additional reason for favoring this approach was that the bulk of the missing data resided on the classroom side of the scale. This is because, in general, DE conditions were described far more completely than their classroom counterparts. This was especially true for media study features because media represents a definitional criterion of DE, whereas it not always present in classrooms. So, in effect many of the relationships expressed in the multiple regression analyses that follow were based on comparisons between a positive value (i.e., either 1 or 2) and 0. Thus, the pedagogical and media study features were recoded using the weighting system shown in Table 3.
Table 3. Pedagogical and media study feature codes and the recodes assigned to them

<table>
<thead>
<tr>
<th>Study Feature Codes and their Meaning</th>
<th>Recodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.  DE more than control group</td>
<td>+2</td>
</tr>
<tr>
<td>2.  DE reported/control group not reported</td>
<td>+1</td>
</tr>
<tr>
<td>3.  DE equal to control group</td>
<td>0</td>
</tr>
<tr>
<td>4.  Control reported/DE not reported</td>
<td>–1</td>
</tr>
<tr>
<td>5.  DE less than control group</td>
<td>–2</td>
</tr>
<tr>
<td>999. Neither DE nor Control Group used explicitly or missing</td>
<td>0</td>
</tr>
</tbody>
</table>

Nine pedagogical study features were: 1) systematic instructional design procedures used; 2) advance course information given to students; 3) opportunity for F2F contact with the teacher; 4) opportunity for F2F contact among student peers; 5) opportunity for mediated communication (e., email, CMC) with the teacher; 6) opportunity for mediated communication among students; 7) student/teacher contact encouraged through activities or course design; 8) student/student contact encouraged through activities or course design; and 9) use of problem-based learning.

The media-related items were: 1) use of two-way audio conferencing; 2) use of two-way videoconferencing; 3) use of computer mediated communication (CMC); 4) use of e-mail 5) use of one-way TV or video or audiotape; 6) use of the Web; 7) use of the telephone; and 8) use of computer-based instruction (CBI).

Data Analysis

Aggregating effect sizes. The weighted effect sizes were aggregated to form an overall weighted mean estimate of the treatment effect (i.e., $g^+$). Thus, more weight was given to findings that were based on larger sample sizes. The significance of the mean effect size was judged by its 95% confidence interval and a z-test. A significantly positive (+) mean effect size indicates that the results favor DE conditions; a significantly negative (–) mean effect size indicates that the results favor traditional classroom-based instruction.

For one study with retention outcomes (Hittleman, 2001) that had extremely large sample sizes (e.g., 1,000,000+), the control sample sizes were reduced to 3,000 with the experimental group’s $N$ reduced proportionally. The treatment $k$ was then proportionally weighted. This procedure was used to avoid overweighting by one study. Outlier analyses were performed using the homogeneity statistics reduction method of Hedges and Olkin (1985) and one achievement.

Testing the homogeneity assumption. In addition, Hedges and Olkin’s (1985) homogeneity procedures were employed in analyzing the effect sizes for each outcome. This statistic, $Q_W$, is an extremely sensitive test of the homogeneity assumption that is evaluated using the sampling distribution of $\chi^2$.

To determine whether the findings for each mean outcome shared a common effect size, the set of effect sizes was tested for homogeneity by the homogeneity statistic ($Q_T$). When all findings share the same population effect size, $Q_T$ has an approximate $\chi^2$ distribution with $k – 1$ degrees of freedom, where $k$ is the number of effect sizes. If the obtained $Q_T$ value is larger than the critical value, the findings are determined to be significantly heterogeneous, meaning that there is more variability in the effect sizes than chance fluctuation would allow. Study features analyses were then performed to identify potential moderating factors.
In the study feature analyses, each coded study feature with sufficient variability was tested through two homogeneity statistics, between-class homogeneity ($Q_B$) and within-class homogeneity ($Q_W$). $Q_B$ tests for homogeneity of effect sizes across classes. It has an approximate $\chi^2$ distribution with $p – 1$ degrees of freedom, where $p$ is the number of classes. If $Q_B$ is greater than the critical value, it indicates a significant difference among the classes of effect sizes. $Q_W$ indicates whether the effect sizes within each class are homogeneous. It has an approximate $\chi^2$ distribution with $m – 1$ degrees of freedom, where $m$ is the number of effect sizes in each class. If $Q_W$ is greater than the critical value, it indicates that the effect sizes within the class are heterogeneous. Data-analyses were conducted using a meta-analysis software, Comprehensive Meta-Analysis™ (Biostat, 2000) and SPSS™ (version 11 for the Macintosh OS X, version 10.2.8).

**Multiple regression modeling of study features.** Weighted multiple regression in SPSS was used to explore the variability in effect sizes and to model the relationships that exist among methodology, pedagogy and media study features. Each effect size was weighted by the inverse of its sampling variance. The equation used in calculating the variance is shown in Equation 3 and the weighting factor in Equation 4 (Hedges & Olkin, 1985, p. 174).

**Equation 3:**
\[
\sigma^2_d = \frac{n_E + n_C}{n_E n_C} + \frac{d^2}{2(n_E + n_C)}
\]

**Equation 4:**
\[
W_i = \frac{1}{\sigma^2_d}
\]

Each set of study features, methodological, pedagogical and media, was entered into weighted multiple regression separately in blocks using $g$ as the dependent variable and $W_i$ as the weight. Methodology, pedagogy and media were entered in different orders to assess the relative contribution ($R^2$ change) of each. Individual methodological, pedagogical and media study features were then assessed to determine their individual contributions to overall variability. To test the significance of individual study features, the individual $\beta$ for each predictor was used but the standard errors were corrected according to Equation 5 (Hedges & Olkin, 1985, p. 174).

**Equation 5:**
\[
SE_{adj} = \frac{SE}{\sqrt{MS_E}}
\]

The 95% confidence interval was corrected using Equation 6 (Hedges & Olkin, 1985, p. 171).

**Equation 6:**
\[
C.I._{adj} = \bar{\sigma} \pm 1.96 \sqrt{\bar{\sigma}}
\]

where, $\bar{\sigma} = (SE_{adj})^2$

The test statistic $z$ was created for testing the null hypothesis that $\bar{\sigma} = 0$ using Equation 7 and $\bar{\sigma}$ was evaluated using $t = 1.96$ (Hedges & Olkin, 1985, p. 172).

**Equation 7:**
\[
z_{\bar{\sigma}} = \frac{\bar{\sigma}}{\sqrt{\bar{\sigma}}}
\]
Results

In total, 232 studies yielding 688 independent effect sizes (i.e., outcomes) were analyzed. This was based on totals of 57,019 students ($k = 321$) with achievement outcomes, 35,365 students ($k = 262$) with attitude outcomes and ($N = 57,916,029$) students ($k = 105$) with retention outcomes. The $N$ reported here for retention was reduced to 3,744,869 to avoid overestimation based on a California study of retention over a number of years. The procedure used in reducing these numbers is described in the section on retention outcomes.

Missing Information

One of the most pervasive finding and difficult problems we encountered in this analysis was the amount of missing information in the research literature. This, of course, was not a problem for the calculation of effect sizes because the availability of appropriate statistical information was a condition of inclusion. However, it was particularly acute in the coding of study features. Table 4 shows a breakdown of missing study feature data over the three outcome measures: achievement, attitudes and retention. Overall, nearly 60% of the potentially codable study features were found to be missing. It is because of this difficulty that we recommend caution in interpreting the results based on study features, including methodological quality. Had the research reports been more complete, we would have been able to offer substantially better-quality advice as to what works and what doesn’t work in DE.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Total Cells</th>
<th># Missing</th>
<th>% Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achievement</td>
<td>13,650</td>
<td>7,726</td>
<td>56.61</td>
</tr>
<tr>
<td>Retention</td>
<td>4,410</td>
<td>2,664</td>
<td>60.41</td>
</tr>
<tr>
<td>Attitude</td>
<td>11,088</td>
<td>5,855</td>
<td>52.80</td>
</tr>
<tr>
<td>Total</td>
<td>29,148</td>
<td>16,246</td>
<td>55.73</td>
</tr>
</tbody>
</table>

Achievement Outcomes

*Total achievement outcomes.* The total number of achievement outcomes was reduced by three outliers, two that exceeded ±3.0 standard deviations from the mean weighted effect size and one whose $Q_W$ was extreme (i.e., > 500). This left 318 achievement outcomes ($N = 54,775$) to be analyzed. Table 5 shows the frequency and percentage of achievement outcomes classified by their date of publication and Table 6 shows their source of publication.

<table>
<thead>
<tr>
<th>Categories of Publication Date</th>
<th>Frequency</th>
<th>Relative Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985 – 1989</td>
<td>27</td>
<td>8.49</td>
</tr>
<tr>
<td>1990 – 1994</td>
<td>91</td>
<td>28.61</td>
</tr>
<tr>
<td>1995 – 1999</td>
<td>108</td>
<td>33.96</td>
</tr>
<tr>
<td>2000 – 2002</td>
<td>92</td>
<td>28.93</td>
</tr>
</tbody>
</table>
Based on these data (Table 5) it is clear that the impetus to conduct research of this kind is not diminishing with time, in spite of calls from prominent voices in the field (e.g., Clark, 1983, 1994) that it should. The Pearson Product Moment correlation between year of publication and Hedges $g$ is $-0.035$ ($df = 316, p < .05$) indicating that there is no systematic relationship between these two variables. In addition, Examination of $g^+$ (Table 6) indicates that there is no substantial bias over the three classes of publication sources upon which these data are based.

Table 7 shows the weighted mean effect size for 318 outcomes. It is essentially zero but the test of homogeneity indicates that wide variability surrounds it. This means that the actual average effect size in the population could range substantially on either side of this value.

The overall distribution of 318 achievement outcomes is shown in Figure 1. It is a symmetrical distribution with a near zero mean, as indicated, a standard deviation of $\pm 0.439$, skewness of 0.203 and kurtosis of 0.752; the distribution is nearly normal. It is clear from the range of effect sizes, from $-1.31$ to $+1.41$ that some applications of DE are far better than classroom instruction and that some are far worse.
Synchronous and asynchronous DE. Achievement outcomes were split using the coded study feature SIMUL, where 1 = synchronous DE, 2 = asynchronous and 3 = unclassified (i.e., missing data). This resulted in 92 synchronous outcomes (N = 8,677), 174 asynchronous outcomes (N = 36,531) and 52 unclassified outcomes (N = 9,567). The mean effect sizes (g+), standard errors, confidence intervals and homogeneity statistics for these three categories are shown in Table 8. The difference in g+ resulting from this split, with synchronous DE significantly negative and asynchronous significantly positive, is dramatic, but both groups remain heterogeneous. Further exploration of the variability in g is required.

Table 8. Weighted mean effect sizes for achievement outcomes for synchronous, asynchronous DE and unclassified DE achievement outcomes

<table>
<thead>
<tr>
<th>Categories of DE</th>
<th>Effect Size</th>
<th>95% Confidence Interval</th>
<th>Homogeneity of ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Hedges’ g+</td>
<td>Standard Error</td>
</tr>
<tr>
<td>Synchronous DE (k = 92) N = 8,677</td>
<td>−0.1022*</td>
<td>0.0236</td>
<td></td>
</tr>
<tr>
<td>Asynchronous DE (k = 174) N = 36,531</td>
<td>0.0527*</td>
<td>0.0121</td>
<td>0.0289</td>
</tr>
<tr>
<td>Unclassified DE (k = 52) N = 9,567</td>
<td>−0.0359</td>
<td>0.0273</td>
<td>−0.0895</td>
</tr>
</tbody>
</table>

*p < .001

Weighted multiple regression. In beginning to explore the variability in g, we conducted WMR with the three blocks of predictors. We were particularly interested in the variance accounted for by each of the blocks—methodology, pedagogy and media—entered in different orders to explore their relative contribution to achievement. Clark and others have argued that: 1) poor
methodological quality tends to confound the effect attributable to features of pedagogy and media; and 2) pedagogy and media, themselves, are confounded in studies of this type. In this analysis, we have attempted to untangle these confounds and to suggest where future researchers and designers of DE applications should expend their energy. WMR was used to assess the relative contributions of these three blocks of predictors. The weighting factor, as described in the Method section, is the inverse of the variance and the dependent variable in all cases was $g$ (Hedges & Olkin, 1985).

We begin with an overall analysis that is followed by a more detailed, albeit more speculative, accounting of the particular study features that account for the more general findings. We entered the three blocks of predictors\(^1\) (e.g., 13 methodological study features) into WMR in different orders: 1) methodology then pedagogy then media; 2) methodology then media then pedagogy; 3) pedagogy then media then methodology; and 4) media then pedagogy then methodology. We did not enter methodology on the second step because this combination seemed counter-intuitive. Table 9 shows the partitioning of variance between ($Q_B$) and within ($Q_W$) on the third step of regression for both synchronous and asynchronous DE outcomes. These statistics are applicable for all of the orders. Notice that $Q_B$ is significant for both DE patterns, and that synchronous DE outcomes are homogeneous (i.e., $Q_W$ is not significant) while asynchronous DE outcomes are not (i.e., $Q_W$ is significant). The critical value of $\chi^2$ at each appropriate degrees of freedom was used to test the significance of each effect.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example.png}
\caption{Tests of between and within-group variation for synchronous DE and asynchronous DE achievement outcomes.}
\end{figure}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
Source & Synchronous DE & & Asynchronous DE & \\
& $SS$ & $df$ & $SS$ & $df$ \\
\hline
$Q_B$ & 111.32* & 26 & 222.41* & 30 \\
$Q_W$ & 66.96 & 65 & 548.84* & 143 \\
Total & 178.29 & 91 & 771.25 & 173 \\
\hline
\end{tabular}
\end{table}

\* $p < .05$

Table 10 provides a comparison of the $R^2$ changes for each of the blocks of predictors. This table reveals some interesting insights into the nature of these predictors, relative to one another. First, note that with one exception each (i.e., third step in both cases), methodology and pedagogy are always significant, no matter which position they are in or whether the outcomes are associated with synchronous or asynchronous DE. Second, note that only when media is entered on the first step is it significant. Overall, this says that the influences of methodology and pedagogy on achievement are more important predictors than media. Third, in line with much of the commentary on the research literature of DE and other media comparison literatures, research

\footnote{We explored another method of entering the three sets of study features in blocks. First, we ran each block separately and saved the unstandardized predicted values ($Y''$). This provided three new composite variables, which were then entered into WMR in different orders, as indicated above. The results were very similar to the ones reported above, with the exception that a clearer distinction emerged between pedagogy and media (i.e., pedagogy was always significant and media was never significant). However, we chose to report the results in the manner described above because it allows a detailed analysis of the contribution of individual study features, whereas the method just described does not. Also, neither synchronous nor asynchronous DE formed a homogeneous set.}
methodology accounts for a substantial proportion of variation in effect size; more for synchronous than for asynchronous DE. One of the difficulties with previous meta-analyses of these literatures is that, at best, methodologically unsound studies were removed a priori, often by fuzzy criteria such as “more than one methodological flaw.” By including studies that range in methodological quality, we have overcome this difficulty to an extent.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>1st Step</th>
<th>2nd Step after Methodology</th>
<th>2nd Step after Pedagogy/Media</th>
<th>3rd Step</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Synchronous DE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methodology</td>
<td>0.490*</td>
<td>—</td>
<td>—</td>
<td>0.250*</td>
</tr>
<tr>
<td>Pedagogy</td>
<td>0.360*</td>
<td>0.101*</td>
<td>0.130**</td>
<td>0.077</td>
</tr>
<tr>
<td>Media</td>
<td>0.245*</td>
<td>0.058</td>
<td>0.015</td>
<td>0.048</td>
</tr>
<tr>
<td><strong>Asynchronous DE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methodology</td>
<td>0.117*</td>
<td>—</td>
<td>—</td>
<td>0.054</td>
</tr>
<tr>
<td>Pedagogy</td>
<td>0.156*</td>
<td>0.107*</td>
<td>0.124*</td>
<td>0.120*</td>
</tr>
<tr>
<td>Media</td>
<td>0.111*</td>
<td>0.051</td>
<td>0.078</td>
<td>0.065</td>
</tr>
</tbody>
</table>

*P < .05 **p = .054 Note. Not all significance tests are based on the same degrees of freedom.

Finally, we calculated the predicted \( g^+ \) that resulted after each step of WMR, with methodology entered first, pedagogy entered second and media entered last. The statistics in Table 11 are the equivalents of the predicted means in \( ANCOVA \) and are the expected values if each of these blocks could have been controlled for in the original experiments and can be compared to the unadjusted \( g^+ \) in row one of the table. Interestingly, \( g^+ \) for synchronous outcomes increases and \( g^+ \) for asynchronous outcomes decreases.

<table>
<thead>
<tr>
<th>Blocks of Predictors</th>
<th>Synchronous DE</th>
<th>Asynchronous DE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( g^+_{predicted} )</td>
<td>( g^+_{predicted} )</td>
</tr>
<tr>
<td>Unadjusted</td>
<td>−0.1022*</td>
<td>0.0527*</td>
</tr>
<tr>
<td>After Methodology</td>
<td>−0.0599</td>
<td>0.0328</td>
</tr>
<tr>
<td>After Pedagogy</td>
<td>−0.0145</td>
<td>0.0238</td>
</tr>
<tr>
<td>After Media</td>
<td>−0.0555</td>
<td>0.0425</td>
</tr>
</tbody>
</table>

* Actual \( g^+ \).

**Study feature analysis.** We now proceed to a more detailed analysis of study features that were outcomes of the previous analysis. A logical way of approaching this is to present the results with methodology on the first step and then both pedagogy and media on separate WMR runs on the second step. This allows for an assessment of pedagogy and media, independently, after variation in methodology has been removed. Therefore, looking back to the previous table, these results derive from pedagogy and media in the second column.
Table 12 shows the individual study feature results for synchronous DE outcomes and Table 13 shows the same results for asynchronous DE outcomes. Shown are the original betas and standard errors along with the adjusted standard errors (see Equation 5), the adjusted confidence intervals (see Equation 6) and the adjusted z-tests (Equation 7), evaluated with $t_{critical} = 1.96$ at a probability of .05.

### Table 12. Significant individual study features for synchronous achievement outcomes.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>SE</th>
<th>SE_adj.</th>
<th>Lower_adj.</th>
<th>Upper_adj.</th>
<th>z-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedagogy Study Features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Face-to-face with instructor</td>
<td>-0.239</td>
<td>0.074</td>
<td>0.07249</td>
<td>-0.3810</td>
<td>-3.29*</td>
</tr>
<tr>
<td>• Face-to-face with Students</td>
<td>0.183</td>
<td>0.073</td>
<td>0.07151</td>
<td>0.0428</td>
<td>2.55*</td>
</tr>
<tr>
<td>Media Study Features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• TV-Video</td>
<td>0.186</td>
<td>0.092</td>
<td>0.08751</td>
<td>0.0144</td>
<td>2.12*</td>
</tr>
<tr>
<td>• Telephone</td>
<td>-0.157</td>
<td>0.065</td>
<td>0.06183</td>
<td>-0.2781</td>
<td>-2.54*</td>
</tr>
</tbody>
</table>

*P < .05, $t_{critical} = 1.96$, $^1MS = 1.166$, $^2MS = 1.042$, $^3MS = 1.105$

### Table 13. Significant individual study features for asynchronous achievement outcomes.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>SE</th>
<th>SE_adj.</th>
<th>Lower_adj.</th>
<th>Upper_adj.</th>
<th>z-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedagogy Study Features</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Advance Info</td>
<td>0.120</td>
<td>0.047</td>
<td>0.024</td>
<td>0.074</td>
<td>5.08*</td>
</tr>
<tr>
<td>• Mediated Comm. Instructor</td>
<td>0.128</td>
<td>0.057</td>
<td>0.029</td>
<td>0.072</td>
<td>4.47*</td>
</tr>
<tr>
<td>• Problem-based Learning</td>
<td>0.280</td>
<td>0.145</td>
<td>0.073</td>
<td>0.137</td>
<td>3.84*</td>
</tr>
<tr>
<td>Media Study Feature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• TV-Video</td>
<td>0.124</td>
<td>0.069</td>
<td>0.034</td>
<td>0.058</td>
<td>3.69*</td>
</tr>
</tbody>
</table>

*P < .05. $t_{critical} = 1.96$, $^1MS = 4.256$, $^2MS = 3.966$, $^3MS = 4.222$

**Demographic study features.** We also coded a set of study features relating to the demographics of students, instructors, subject matter and reasons for offering DE. Table 14 contains the three that yielded enough outcomes to warrant analysis. DE achievement effects were large when: a) the efficient delivery or cost was a reason for offering DE courses ($g^+ = 0.1639$); b) for students in grades k-12 ($g^+ = 0.2714$); and for military and business applications ($g^+ = .1777$).
<table>
<thead>
<tr>
<th>Study Features</th>
<th>Hedges’ $g^+$</th>
<th>z-value</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reasons for offering DE courses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Access to expertise ($k = 48$)</td>
<td>–0.0821</td>
<td>–2.93</td>
<td>&lt; .005</td>
</tr>
<tr>
<td>• Efficient delivery or cost ($k = 22$)</td>
<td>0.1639</td>
<td>3.55</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>• Multiple purposes ($k = 22$)</td>
<td>0.1557</td>
<td>2.84</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Type of students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• K-12 ($k = 16$)</td>
<td>0.2714</td>
<td>4.88</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>• Undergraduate ($k = 123$)</td>
<td>0.0138</td>
<td>0.73</td>
<td>&gt; .05</td>
</tr>
<tr>
<td>• Graduate ($k = 26$)</td>
<td>–0.0984</td>
<td>–1.96</td>
<td>= .05</td>
</tr>
<tr>
<td>• Military ($k = 10$)</td>
<td>0.4831</td>
<td>7.20</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Subject matter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Math, science and engineering ($k = 67$)</td>
<td>–0.1026</td>
<td>–3.94</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>• Computer science/computer applications ($k = 13$)</td>
<td>0.1706</td>
<td>3.01</td>
<td>&lt; .005</td>
</tr>
<tr>
<td>• Military and business/industry ($k = 50$)</td>
<td>0.1777</td>
<td>5.72</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

**Attitude Outcomes**

*Synchronous and asynchronous outcomes.* We found various forms of attitude measures in the literature that could be classified into four categories: attitudes towards technology; attitudes towards subject matter; attitudes towards instructor; and attitudes towards course. We also have a fairly large set of measures ($k = 90$) that could not be classified into a single set, and therefore were labeled “Other Attitude Measures.” We chose not to include “Other Attitudes” in the same analysis where the type of measure was known. Therefore, the total number of attitude outcomes was reduced from 262 to 172. This number was further reduced when missing data prevented us from categorizing outcomes as either synchronous or asynchronous. Before analysis, one extremely high outlier was removed. This left 154 outcomes to be analyzed.

We split the sample in synchronous and asynchronous DE, in the same manner as for achievement, and found essentially the same overall dichotomy. Table 15 shows these results along with the results of 154 combined attitudes. While all of the weighted mean effect sizes are negative, notice the contrast between synchronous and asynchronous outcomes. The average effect size for synchronous outcomes is significant, while the average effect size for asynchronous is not. Furthermore, there is high variability among effect sizes, even after the split. Figure 3 provides a graphic depiction of overall variability in attitude outcomes for 154 outcomes and shows that they range from +2.41 to –2.38. There are circumstances where DE student reactions are extremely positive and others where reactions are quite negative, relative to classroom instruction.
Table 15. Weighted mean effect sizes for combined attitude outcomes and for synchronous and asynchronous DE attitude outcomes

<table>
<thead>
<tr>
<th>Categories of DE</th>
<th>Effect Size</th>
<th>95% Confidence Interval</th>
<th>Homogeneity of ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hedges’ $g^+$</td>
<td>Standard Error</td>
<td>Lower Limit</td>
</tr>
<tr>
<td>Combined (not including “other attitudes” ($k = 154$)</td>
<td>$N = 21,047$</td>
<td>$-0.0812^*$</td>
<td>0.0146</td>
</tr>
<tr>
<td>Synchronous DE ($k = 83$)</td>
<td>$N = 9,483$</td>
<td>$-0.1846^*$</td>
<td>0.0222</td>
</tr>
<tr>
<td>Asynchronous DE ($k = 71$)</td>
<td>$N = 11,624$</td>
<td>$-0.0034$</td>
<td>0.0193</td>
</tr>
</tbody>
</table>

* $p < .001$

Figure 3. Distribution of 154 effect sizes for combined Attitude outcomes.

*Weighted multiple regression.* Given the wide variability in attitude outcomes, a WMR analysis was conducted in a manner that was similar to the one done with achievement data. The within and between-groups tests of significance are shown in Table 16. $Q_w$ is significant for synchronous and asynchronous DE outcomes indicating heterogeneity for these groups.
Figure 16. Tests of between and within-group variation for synchronous DE and asynchronous DE for attitude outcomes

<table>
<thead>
<tr>
<th>Source</th>
<th>Synchronous DE</th>
<th>Asynchronous DE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SS</td>
<td>df</td>
</tr>
<tr>
<td>$Q_B$</td>
<td>260.68*</td>
<td>22</td>
</tr>
<tr>
<td>$Q_W$</td>
<td>135.35*</td>
<td>60</td>
</tr>
<tr>
<td>Total</td>
<td>396.025</td>
<td>82</td>
</tr>
</tbody>
</table>

* $p < .05$

We examined the $R^2$ change for three blocks of predictors, methodology, pedagogy and media for attitudes in different orders, in the same way as we did for achievement outcomes. Table 17 is a comparison of $R^2$ change for blocks of study features entered in different orders in WMR. The results do not do as clearly favor methodology and pedagogy and the diminished role of media as they did for achievement. In fact, these results indicate a more complex relationship among the three blocks of predictors. For one thing, there are more differences here for the three blocks of predictors between synchronous and asynchronous DE. As with achievement, methodology still accounts for more variation in synchronous DE than asynchronous. While pedagogy is somewhat suppressed for synchronous DE, it emerges as important for asynchronous DE. On the other hand, media appears to be more important in synchronous DE than in asynchronous DE.

Table 17. Comparison of $R^2$ change for blocks of study feature predictors entered on different steps of WMR for attitude outcomes

<table>
<thead>
<tr>
<th>Predictors</th>
<th>1st Step</th>
<th>2nd Step after Methodology</th>
<th>2nd Step after Pedagogy/Media</th>
<th>3rd Step</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Synchronous DE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methodology</td>
<td>0.471*</td>
<td>—</td>
<td>—</td>
<td>0.421*</td>
</tr>
<tr>
<td>Pedagogy</td>
<td>0.128</td>
<td>0.138*</td>
<td>0.101</td>
<td>0.120*</td>
</tr>
<tr>
<td>Media</td>
<td>0.136*</td>
<td>0.067*</td>
<td>0.109*</td>
<td>0.049</td>
</tr>
<tr>
<td><strong>Asynchronous DE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methodology</td>
<td>0.218*</td>
<td>—</td>
<td>—</td>
<td>0.157</td>
</tr>
<tr>
<td>Pedagogy</td>
<td>0.253*</td>
<td>0.215*</td>
<td>0.133</td>
<td>0.076</td>
</tr>
<tr>
<td>Media</td>
<td>0.241*</td>
<td>0.236*</td>
<td>0.121</td>
<td>0.097</td>
</tr>
</tbody>
</table>

* $P < .05$  **p = .057** Note. Not all significance tests are based on the same degrees of freedom.

Table 18 shows the predicted $g+$ results before WMR and after methodology, pedagogy and media were entered. Notice that synchronous DE becomes more negative (i.e., favoring the classroom condition) and asynchronous DE moves from an initial negative sign to a positive sign after methodology is entered, and thereafter.
Table 18. Predicted values of $g^+$ at each level of methodology, pedagogy and media for attitude outcomes

<table>
<thead>
<tr>
<th>Blocks of Predictors</th>
<th>Synchronous DE $g^+_{predicted}$</th>
<th>Asynchronous DE $g^+_{predicted}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted</td>
<td>–0.1846*</td>
<td>–0.0034*</td>
</tr>
<tr>
<td>After Methodology</td>
<td>–0.1707</td>
<td>0.0243</td>
</tr>
<tr>
<td>After Pedagogy</td>
<td>–0.1702</td>
<td>0.0212</td>
</tr>
<tr>
<td>After Media</td>
<td>–0.2031</td>
<td>0.0237</td>
</tr>
</tbody>
</table>

* Actual $g^+$.

**Study feature analysis.** Individual study features were assessed after WMR in a manner similar to that for achievement outcomes and the significant results are shown in Table 22 for synchronous DE outcomes and in Table 19 for asynchronous DE outcomes. Again, regression information from methodology on the first step, pedagogy on the second step (i.e., after methodology) and media on the second step (i.e., after methodology) is presented. The adjustments indicated to standard errors, confidence intervals and the $z$-test were performed according to the equations in the Method section.

Table 19. Assessment of individual study features in WMR for synchronous attitude outcomes

<table>
<thead>
<tr>
<th>Predictors</th>
<th>$b$</th>
<th>SE</th>
<th>$SE_{adj.}$</th>
<th>Confidence Intervals</th>
<th>$z$-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower$_{adj.}$</td>
<td></td>
</tr>
<tr>
<td>Pedagogy Study Features (Step 2 after methodology)</td>
<td></td>
<td></td>
<td></td>
<td>Upper$_{adj.}$</td>
<td></td>
</tr>
<tr>
<td>• Advance Info.</td>
<td>–0.281</td>
<td>0.138</td>
<td>0.089$^2$</td>
<td>–0.455</td>
<td>–0.107</td>
</tr>
<tr>
<td>• Systematic ID</td>
<td>1.162</td>
<td>0.385</td>
<td>0.248</td>
<td>0.677</td>
<td>1.647</td>
</tr>
<tr>
<td>• Mediated comm. with instructor</td>
<td>0.061</td>
<td>0.255</td>
<td>0.164</td>
<td>0.279</td>
<td>0.923</td>
</tr>
<tr>
<td>• Instructor/student contact encouraged</td>
<td>0.314</td>
<td>0.135</td>
<td>0.087</td>
<td>0.144</td>
<td>0.484</td>
</tr>
<tr>
<td>Media Study Features (Step 2 after methodology)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• TV-Video</td>
<td>–0.351</td>
<td>0.126</td>
<td>0.076$^3$</td>
<td>–0.500</td>
<td>–0.202</td>
</tr>
<tr>
<td>• Use of telephone</td>
<td>0.262</td>
<td>0.103</td>
<td>0.062</td>
<td>0.140</td>
<td>0.384</td>
</tr>
</tbody>
</table>

*p < .05. $t_{critical} = 1.96$, $^1MS = 2.949$, $^2MS = 2.416$, $^3MS = 2.73$
Table 20. Assessment of individual study features in WMR for asynchronous attitude outcomes

<table>
<thead>
<tr>
<th>Predictors</th>
<th>SE</th>
<th>SE_{adj.}</th>
<th>Lower_{adj.}</th>
<th>Upper_{adj.}</th>
<th>z-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedagogy Study Feature (Step 2 after methodology)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Problem-based learning</td>
<td>0.403</td>
<td>0.180</td>
<td>0.209</td>
<td>0.597</td>
<td>4.07*</td>
</tr>
<tr>
<td>Media Study Features (Step 2 after methodology)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• CMC</td>
<td>0.272</td>
<td>0.108</td>
<td>0.151</td>
<td>0.373</td>
<td>4.41*</td>
</tr>
<tr>
<td>• CBI</td>
<td>0.392</td>
<td>0.150</td>
<td>0.224</td>
<td>0.560</td>
<td>4.57*</td>
</tr>
<tr>
<td>• Web</td>
<td>-0.168</td>
<td>0.191</td>
<td>-0.270</td>
<td>-0.066</td>
<td>-3.23*</td>
</tr>
</tbody>
</table>

*p < .05. \( t_{critical} = 1.96 \), \(^1 MS = 3.934\), \(^2 MS = 3.307\), \(^3 MS = 3.063\)

**Retention Outcomes**

Retention is defined here as the opposite of dropout. Since the drop-out rate in DE has been perceived as a problem since its inception as reflected in much of the earlier literature, we wondered if this situation had changed since 1985, with the inception of newer forms of communication. We found several studies of statewide data (i.e., California) that compared DE to classroom conditions, where the sample size was in the millions. To correct for the extreme effects of these huge (\( N = 57,916,029 \)), but anomalous studies, we truncated the sample sizes of the classroom condition to 3,000 and proportionately reduced the DE condition to create a better balance with other studies (\( N = 3,735,050 \)). Otherwise, these effect sizes would have dominated the average effects, unduly skewing it in favor of the large samples. Table 21 shows the results of this analysis and the results of the split between synchronous and asynchronous DE conditions. Note that none of the large-sample studies had been coded as either synchronous or asynchronous and so, while the number of effects is fairly representative of the total, the number of students is not. In spite of this, the results of the synchronous/asynchronous split seem to reflect the average for all studies. Caution in the interpretation of the mean effect size for synchronous DE should be exercised because of the low number of outcomes associated with it.

One note of interest is that the Pearson Product-Moment Correlation between “year of publication” and \( g \) is 0.015 (\( df = 68 \), \( p > .05 \)), suggesting that there is no systematic increase or decrease in the differential retention rate over time. This situation was somewhat different for synchronous (\( r = -0.27, df = 14, p > .05 \)) and asynchronous retention outcomes (\( r = 0.011, df = 51, p > .05 \)) calculated separately, although neither reached significance. Had the synchronous correlation been significant, it would have indicated a decreasing differential (i.e., the two conditions becoming more the same) over time between classroom and DE applications in terms of retention.
Table 21. Weighted mean effect sizes for retention outcomes for synchronous and asynchronous DE

<table>
<thead>
<tr>
<th>Outcome Type</th>
<th>Effect Size</th>
<th>95% Confidence Interval</th>
<th>Homogeneity of ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hedges’ g+</td>
<td>Standard Error</td>
<td>Lower Limit</td>
</tr>
<tr>
<td>Overall Retention</td>
<td>-0.0573*</td>
<td>0.0065</td>
<td>-0.0700</td>
</tr>
<tr>
<td>(k = 103) N = 3,735,050</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synchronous DE</td>
<td>0.0051</td>
<td>0.0341</td>
<td>-0.0617</td>
</tr>
<tr>
<td>(k = 17) N = 3,604</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asynchronous DE</td>
<td>-0.0933*</td>
<td>0.0211</td>
<td>-0.1347</td>
</tr>
<tr>
<td>(k = 53) N = 10,435</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .05

Figure 4. Distribution of effect size for 70 retention outcomes.

When WMR was performed on synchronous and asynchronous retention outcomes, neither the results of methodology, pedagogy or media were significant. Therefore, no regression outcomes are presented.

Summary of Results

1. There is a very small and significant effect favoring DE conditions (g+ = 0.0128) on overall achievement outcomes (k = 318). However the variability surrounding this mean is wide and significant.

2. When outcomes were split between synchronous and asynchronous DE achievement outcomes, a small positive but non-significant effect (g+ = 0.0527) occurred for asynchronous DE and a larger negative significant effect occurred for asynchronous DE.
Variability remained wide and significantly heterogeneous for each group.

3. WMR revealed that together Methodology, Pedagogy, and Media accounted for 62.4% of variation in synchronous DE achievement outcomes and 28.8% of variability in asynchronous DE outcomes.

4. When $R^2$ change was examined, entered into WMR in different orders, methodology and pedagogy were almost always significant, whereas media was only significant when it was entered on the first step. This was true for both synchronous DE and asynchronous DE outcomes. Individual significant study feature outcomes are summarized in Table 22.

5. There is a small negative but significant effect in favor of classroom instruction ($g^+ = -0.0812$), on overall attitude outcomes. Again, the variability around this mean is significantly heterogeneous.

6. There are differences in effect sizes for synchronous DE ($g^+ = -0.1846$) and asynchronous DE ($g^+ = -0.0034$). Both favor classroom instruction but the average effect size for asynchronous DE is significant and it is not for asynchronous DE. Individual significant study feature outcomes are summarized in Table 22.

7. WMR revealed that together Methodology, Pedagogy, and Media accounted for 65.8% of variability in synchronous DE attitude outcomes and 53.1% of variability in asynchronous DE outcomes.

8. $R^2$ change analysis of the type described above revealed varying patterns of variance accounted for methodology, pedagogy, and media. It appears that these three sets of variables are related in a more complex way than they are for achievement outcomes.

9. There is a very small but significant effect, in favor of classroom instruction ($g^+ = -0.0573$) on retention outcomes. There is a very small but positive effect for synchronous DE that is not significant ($g^+ = 0.0051$) and a larger negative effect ($g^+ = -0.0933$) for asynchronous DE.

10. There is extremely wide variability in effect size on all measures, and we have yet to find study features that form homogeneous subsets, including the distinction between synchronous and asynchronous DE (with the one exception of synchronous DE on achievement). This suggests that DE works extremely well sometimes and extremely poorly sometimes, even when all coded study features are accounted for.

11. Since the variation in effect size accounted for by methodology is fairly substantial (generally speaking, more substantial for synchronous than asynchronous DE), and often more than pedagogy and media combined, methodological weakness was considered an important deterrent to forming clear recommendations to practitioners and policymakers.

12. Another measure of the quality of the literature, amount of data available, suggests that the literature is very weak in design features that would improve the interpretability of the results. Over half (55.73%) of the codable study features (including methodological features) were missing.

13. Even though the literature is large, it is difficult to draw firm conclusions on what works and doesn’t work in DE, except to say that the distinction between synchronous and asynchronous forms of DE does moderate effect sizes in terms of both achievement and attitudes. Concise statements of outcomes based on study feature analysis (Table 22)
should proceed with caution and remain speculative because of relatively large amount of missing data relating to them.

Table 22 is a summary of the significant study features that resulted from WMR.

Table 22. Summary of study features that significantly predict achievement, attitude and retention outcome measures

<table>
<thead>
<tr>
<th>Synchronous DE</th>
<th>Asynchronous DE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Favor Classroom Instruction (−)</strong></td>
<td><strong>Favor Distance Education (+)</strong></td>
</tr>
<tr>
<td><strong>Achievement</strong></td>
<td><strong>Achievement</strong></td>
</tr>
<tr>
<td>• Face-to-face meetings with the instructor</td>
<td>• Face-to-face contact with other students</td>
</tr>
<tr>
<td>• Use of the telephone to contact instructor</td>
<td>• Use of one-way TV-video</td>
</tr>
<tr>
<td><strong>Attitudes</strong></td>
<td><strong>Attitudes</strong></td>
</tr>
<tr>
<td>• Opportunity for face-to-face contact with other students</td>
<td>• Use of systematic ID</td>
</tr>
<tr>
<td>• Use of one-way TV-video</td>
<td>• Opportunity for mediated communication with the instructor</td>
</tr>
<tr>
<td><strong>Retention</strong></td>
<td><strong>Retention</strong></td>
</tr>
<tr>
<td>• No significant predictors</td>
<td>• No significant predictors</td>
</tr>
</tbody>
</table>

| **Favor Classroom Instruction (−)** | **Favor Distance Education (+)** |
| **Achievement** | **Achievement** |
| • No significant predictors | • Use of problem-based learning strategies |
| **Attitudes** | **Attitudes** |
| • Use of the Web | • Opportunity for mediated communication with the instructor |
| **Retention** | **Retention** |
| • No significant Predictors | • Advance information given to students |
| | • Use of one-way TV-video |
| | • Use of problem-based learning strategies |
| | • Use of computer-mediated communication |
| | • Use of computer-based instruction |
| | • No significant predictors |
Discussion

Overall Findings
The most important outcome of the overall analysis of effect size relates to the wide variability in outcomes for all three primary measures. While the average effect of DE was near zero, there is a tremendous range of effect sizes ($g$) in achievement outcomes—from +1.41 to −1.31. There are instances where the DE group outperformed the traditional instruction group by more than 50%. And there are instances where the opposite occurred—the traditional instructional group outperformed the DE group by 48% or more. This is likewise for overall attitude and retention outcomes.

None of the measures is homogeneous, so interpreting means as if they are true representations of population values is risky (Hedges & Olkin, 1985). It is simply incorrect to say that DE is better than, worse than or even equal to classroom instruction on the basis of mean effect sizes and heterogeneity. This wide variability means that a substantial proportion of DE applications provide better achievement results, are viewed more positively and have higher retention rates than their classroom counterparts. On the other hand, a substantial proportion of DE applications are far worse than classroom instruction on all three measures.

The mistake that a number of previous reviews have made, from early narrative reviews (e.g., Moore and Thompson, 1990) to more recent reviews (e.g., Russell, 1999), is to declare that DE and classroom instruction are equal without examining the variability surrounding their difference. Wide and unexplained variability precludes any such simplistic conclusion. An assessment of the literature of this sort can only be made through a meta-analysis that provides a comprehensive representation of the literature, the application of rigorously applied inclusion/exclusion criteria and an analysis of variability around mean effect sizes. One further note, the overall retention outcomes appear to indicate that the substantial degree of retention differential between classroom and DE conditions, noted in many studies of student persistence, is still present in these studies.

Quality of the DE Literature
In the last few years, a number of commentators (Anglin & Morrison, 2000; Diaz, 2000; Perraton, 2000; Saba, 2000; Phipps & Merisotis, 1999) have decried the quality of the DE research literature. One of the main purposes of this meta-analysis was to estimate the extent of these claims and to examine the depth of the research literature in terms of its completeness. This discussion begins with that assessment, because both the quality of studies and the depth of reporting impinge upon all other aspects of the analysis.

One whole section of the codebook (13 items) deals with methodological aspects of the studies that were reviewed. Our intent was not to exclude studies that had methodological weaknesses, like lack of random assignment or non-equivalent materials, but to code these features and examine how they affect the conclusions that can be drawn from the studies. However, the quality of reporting in the literature that we examined affects the quality of the methodological assessment, since missing aspects of design, control, measurement, equivalence of conditions, etc. influences the quality of the methodological assessment.

Information available in the literature. Overall, we found the literature severely wanting in terms of depth of reporting. Nearly 60% of the codable study features, including methodological features, were coded as missing. This means that for outcomes that met our inclusion criteria and for which we could calculate an effect size, we were only able to derive a 40% estimate of the
study features on the effect sizes. The most persistent problem was the reporting of characteristics of the comparison condition (i.e., classroom instruction). Often, authors went to extraordinary lengths to describe the DE condition, only to say that it was being compared to a “classroom condition.” If one cannot discern what a DE condition is being compared to, it is very difficult to come to any conclusion as to what an effect size characterizing their difference means. This was not just a problem in reports and conference papers that are often not reviewed or reviewed only at a cursory level; it was true of journal articles and dissertations, as well, which are presumably reviewed by panels of peers or committees of academics. This speaks not only to the quality of the peer review process of journals but to the quality and rigor of training that future researchers in our field are receiving. However, an analysis of publication source did not reveal any substantial bias in mean effect size among the types of literature that are represented in these data (i.e., achievement data only).

There are some interesting statistics associated with year of publication that bear noting. In spite of calls from the field to end the form of classroom comparative studies investigated here (e.g., Clark, 1983, 1994), their frequency actually appears to have been increasing since 1985. As indicated in the Results, there appears to be no systematic relationship between “year of publication” and effect size.

Methodological quality of the literature. Field experiments investigating educational practices are characteristically weak because they are so often conducted in circumstances where the opportunities to control for rival explanations of research hypotheses are minimal. Therefore, they are typically higher in external validity than in internal validity. The What Works Clearinghouse (Valentine & Cooper, 2003) uses a four-axis model of research methodology to judge the quality of a research study: internal validity, external validity, measurement validity and measurement reliability. Our 13 coded study features relating to methodology focused more on internal validity than the other three. Ten items rated aspects of internal validity in terms of the equality or inequality of comparison groups, no direct assessment of external validity was made, one feature assessed the quality of the outcome measure used, another assessed the quality of the publication source and another rated the quality of the statistical information used in calculating effect sizes (i.e., calculated or estimated).

Remembering that many codable aspects of methodological quality were unavailable because of missing information, we tried to characterize the quality of studies in terms of research design and the degree of control for confounding. We chose to enter the 13 methodological study features into weighted multiple regression as a way of: 1) assessing methodology independently and in relation to other blocks of study features, and 2) assessing other study features after variation due to methodology was removed. Overall, we found methodology to account for a substantial proportion of overall variation in the effect sizes for achievement and attitude measures. This was moderated somewhat when outcomes were split between synchronous and asynchronous DE patterns. Typically, more methodological variation was accounted for in synchronous DE than in asynchronous DE.

Our recoding scheme emphasized the difference between methodological strengths and methodological weaknesses, with missing data considered neutral. In a strong experimental literature, with little missing data, strong measures and adequate control over confounding, the variance accounted for by methodology would have been minimal. In the most extreme case, zero variability would be attributable to methodology. As previously indicated, this was not the case, suggesting that the dual contributing factors of experimental and methodological inadequacies and missing information weaken this DE research literature. However, this fact
does not mitigate entirely against exploring these data in an effort to learn more about the characteristics of DE and the relative contributions of various factors to its success or failure, relative to classroom instruction.

**Synchronous and Asynchronous DE**

After assessing overall outcomes for the three measures, we split the samples into the two different forms of DE noted in the literature, synchronous DE and asynchronous DE. Synchronous DE is defined by the time- and place-dependent nature of classroom instruction proceeding in synchronization with a DE classroom, located in a remote location and connected by videoconferencing, audio-conferencing media or both. Asynchronous DE conditions were run independently of their classroom comparison conditions. While a few asynchronous applications actually used synchronous media among themselves, they were not bound by time and place to the classroom comparison condition. The current use of the word asynchronous often refers to the lag-time in communication that distinguishes, for instance, email from a “chat room”; our definition does not disqualify some synchronous communication between students and instructors and students and other students.

The results of this split yielded substantially different outcomes for the two forms of DE on all three measures. In the case of achievement, synchronous outcomes favored the classroom condition, ranging from +0.97 to –1.14 (this is the only homogeneous subset), while asynchronous outcomes favored the DE condition, ranging from +1.41 to –1.31. While both mean effect sizes for attitudes were negative, the differences were dramatically different for synchronous and asynchronous DE, favoring classroom instruction by nearly 0.20 SD. The split for retention outcomes yielded the opposite outcome. Dropout was substantially higher in asynchronous DE, compared with synchronous DE.

It is possible that these three results can be explained in the same terms by examining the conditions under which students learn and develop attitudes in these two patterns, as well as make decisions to persist or drop out. Looked at in one way, synchronous DE is a poorer-quality replication of classroom instruction; there is neither the flexibility of scheduling and place of learning nor the individual attention that exists in many applications of asynchronous DE, and there is the question of the effectiveness of “face-to-face” instruction conducted through a teleconferencing medium. Although we were unable to ascertain much about teaching style from the literature, there may be a tendency for synchronous DE instructors to engage in lecture-based, instructor-oriented strategies that may not translate well to mediated classrooms at a distance (Verduin & Clark, 1991). Even employing effective questioning strategies may be problematic under these circumstances. In fact, there have been calls in the literature of synchronous DE for instructors to adopt more constructivist teaching practices (Beaudoin, 1990; Dillon & Walsh, 1992; Gehlauf, Shatz, & Frye, 1991). According to Bates (1997) asynchronous DE, by contrast, can more effectively provide interpersonal interaction and support two-way communication between instructors and students and among students, thereby producing a better approximation to a learner-centered environment. These two sides of the DE coin may help explain the differential achievement and attitude results.

Work carried out by Chickering and Gamson (1987) offers an interesting framework with which addressing the question of teaching in DE environments. Based on 50 years of higher education research, they produced a list of seven principles of good teaching practices in face-to-face courses. Graham, Cagiltay, Craner, Lim and Duff (2000) used these same seven principles to assess whether these skills transfer to online teaching environments. Their general findings, echoed by the work of Schoenfeld-Tacher and Persichitte (2000) and Spector (2001), indicate...
that DE teachers typically require different sets of technical and pedagogical competencies to engage in superior teaching practices, although Kanuka, Collett & Caswell (2003) claim that this transition can be made fairly easily by experienced instructors. Presumably, this applies to both synchronous and asynchronous DE, but because synchronous DE is more like classroom instruction, and is performed in view of a live classroom as well as a mediated one, it is possible that adopting new and more appropriate teaching methods is not as critical and pressing as it is in asynchronous DE.

If achievement is better and attitudes are more positive in asynchronous DE than in synchronous DE, why is its retention rate lower? First of all, based on the literature, it is not surprising that there is greater dropout in DE courses than in traditional classroom-based courses (Kember, 1996). The literature has said this for years. However, this does not fully answer the question about synchronous and asynchronous DE. Part of the answer is that achievement and attitude measurement are independent of retention since they do not include data from students who dropped out before the course ended. A second part of the answer may lie, again, in differences in the different conditions that exist in synchronous and asynchronous DE. As previously noted, synchronous DE is more like classroom instruction than asynchronous DE. Students meet together, in a particular place, at a particular time. They are a group, just like the classroom students. The difference is that they are remote from the instructor. Students working in asynchronous DE conditions do not typically meet in groups, although they may have face-to-face and/or synchronous mediated contact with the instructor and other students. Group affiliation and social pressure, then, may partially explain this effect. Other explanations may derive from models of persistence such as Kember’s (1996), which stress factors such as entry characteristics, social integration, external attribution and academic integration.

Only a small percentage of the findings for synchronous DE are based on K-12 learners. We speculate that for younger learners the structure of synchronous DE may be better suited to their academic schedules and need for organization, spontaneous guidance and feedback. Furthermore, we have concerns about the nature of appropriate comparisons. For example, how does synchronous DE compare to home schooling or the provision of specialized content by a non-expert (e.g., in rural and remote communities)?

This is an even more general concern that goes beyond synchronicity of DE delivery and addresses the question of access to education and the appropriate nature of the comparison condition. When is it appropriate for DE to be compared to traditional instruction, other alternative delivery methods or a no instruction control group? In the latter case, this may be the choice with which a substantial number of learners are faced and represents one purpose of DE—to provide learning opportunities when no others exist. In such circumstances, issues of instructional quality, attitudes and retention may be secondary to issues of whether assessment and outcome standards—ensuring rigorous learning objectives—are maintained.

Media vs. Pedagogy: Resolving the Debate?
Is technology transparent or is it transformative? Do the most effective forms of DE take unique advantage of communication and multimedia technologies in ways absent from “traditional” classroom instruction? If so, why are these absent from classroom instruction? For example, how much does the distance-learning context provide the requisite incentive for learners to use the technological features apparent in some media-rich DE applications? Alternatively, can effective pedagogy exist independently of the affordances and restrictions of DE? Can, for example, clarity, expressiveness and instructional feedback be provided regardless of their medium of
delivery and independently of the separation of space and time? Finally, how can we begin to explore these issues independently of concerns about methodological quality and completeness?

The nature of the DE research literature, in which research methodology, pedagogy and media are all present and intertwined, gave us an opportunity to examine their relative contributions to achievement, attitude and retention outcomes and to further explore the wide variability that still existed after studies were split into synchronous and asynchronous DE. We settled on an approach to weighted multiple regression (WMR) in which blocks of these recoded study features were entered in different orders and assessed the $R^2$ change that resulted from their various positions in the regression models. With the exception of retention, which did not achieve statistical significance for either type of DE, the overall percentage of variance accounted for by these blocks ranged from 29% to 66% for achievement and attitude. However, only one homogeneous set was found: achievement outcomes for synchronous DE.

Methodology. In the design of original experimental research, the more the extraneous differences between treatment and control can be minimized, the stronger the causal assertion. However, in a meta-analysis, actual control cannot be applied to the studies under scrutiny, so that the best that can be done is to estimate the methodological strength or weakness of the research literature. The first thing we found is that methodology is a good predictor of achievement and attitude effect sizes, but a better predictor in synchronous DE studies (49% and 47%, respectively) than in asynchronous DE studies (12% and 22%). Second, we found that methodology is a strong predictor of achievement and attitude effect size, whether entered on the first or the third step of WMR for synchronous DE, but not for asynchronous DE. Because of the way methodology was recoded, this means that studies of asynchronous DE are of higher quality than studies of synchronous DE.

Pedagogy and media. Clark (1983, 1994) has argued vociferously that media and technology, used in educational practice, have no effect on learning. Instead, it is the characteristics of instructional design, such as the instructional strategies that are used, the feedback that is provided and the degree of learner engagement, that create the conditions within which purposive learning will occur. In general, we found this to be the case. Characteristics of pedagogy tended to take precedence over media, no matter on which step in WMR they were entered. This is especially true for achievement outcomes; the relationship for attitudes is a little more complex. Does this mean that media isn’t important? No, it can’t mean that because media is a requirement for DE to exist in the first place. It does mean, however, that instructional practices, independent of the medium, are critical to all forms of educational practice, including, and possibly especially, DE. This seems almost too axiomatic to state, and yet in the literature of DE there is an exaggerated emphasis on the medium du jour. As Richard Clark recently explained (personal communication, April and October, 2003), it was the tendency of educational technologists to become enamored with “the toys of technology” that led to his original thesis and his continued insistence that media are of little concern compared with the myriad elements of sound instructional practice. There is a now old instructional design adage that goes something like this: “a medium should be selected in the service of instructional practices, not the other way around.” We would encourage all practitioners and policymakers, bent on developing and delivering quality DE, whether on the Internet or through synchronous teleconferencing, to heed this advice.

Considerations for Practice
Before moving on to a discussion of individual study features, there are two issues that need reiteration. First, interpretation of individual predictors in WMR, when overall results are
heterogeneous, must proceed with caution (Hedges & Olkin, 1996). Second, some of the individual study feature results are based on a fairly small number of actual outcomes and therefore must be taken as speculative.

Specific considerations. Unfortunately, we are unable to offer any recipes for the design and development of quality DE. Missing information in the research literature, we suspect, is largely responsible for this. However, we are able to speak in broad terms about some of the things that matter in synchronous and asynchronous DE applications:

• Attention to quality course design should take precedence over attention to the characteristics of media. This presumably includes what the instructor does as well as what the student does, although we only limited direct evidence of the either. However, the appearance of “use of systematic instructional design” as a predictor of attitude outcomes implicates instructors and designers of asynchronous DE conditions.

• Active learning (e.g., PBL) that includes (or induces) some collaboration among students appears to foster better achievement and attitude outcomes in asynchronous DE.

• Opportunities for communication, both face-to-face and through mediation, appear to benefit students in synchronous and asynchronous DE.

• “Supplementary one-way video materials” and “use of computer-based instruction” were also found to help promote better achievement and attitude outcomes in synchronous and asynchronous DE.

• In asynchronous DE, Media that support interactivity (i.e., CMC and the telephone) appear to facilitate better attitudes and “providing advance course information” benefits achievement outcomes.

The similarities between the results of achievement and attitude across synchronous and asynchronous DE are both strikingly similar and strikingly different. For instance, for asynchronous DE, problem-based learning (PBL) appears as a strong predictor in favor of the DE condition. Although this is one of the study features with relatively few instances, we speculate that it is the collaborative, learner-oriented aspect of this instructional strategy that accounts for better achievement and more positive attitudes. Judging from reviews in the medical education literature (e.g., Albanese & Mitchell, 1993; Colliver, 1999), where 30 years of studies have been performed with PBL, this instructional strategy represents a useful mechanism for engaging students, teaching problem-solving and developing collaborative working skills. We suggest that more research be pursued in this area, along the lines of the study conducted by Bernard and Lundgren-Cayrol (2001), which explored ways that PBL is best suited for promoting DE learning achievement. Bernard, Rojo de Rubalcava & St. Pierre (2000) describe other ways that PBL might be linked to collaborative learning in online learning environments.

Among the other pedagogical study features is a group of features that relate to both face-to-face and mediated contact with the instructor in a course and among student peers. We also found that “encouragement of contact (either face-to-face or mediated)” predicted outcomes for both synchronous and asynchronous DE, when achievement and attitudes were examined jointly. This suggests that DE should not be a solitary experience as it often was in the era of correspondence education. Instructionally relevant contact with instructors and peers is not only desirable, it is probably necessary for creating learning environments that lead to desirable achievement gains and general satisfaction with DE. This is not a particular revelation, but it is an important aspect of quality course design that should not be neglected or compromised.
One of the surprising aspects of this analysis is that the mechanisms of mediated communication (e.g., email) did not figure more prominently as predictors of learning or attitude outcomes. Computer-mediated communication did arise as a significant predictor of attitude outcomes, but a rather traditional medium, the telephone, also contributed to the media equation. In addition, non-interactive one-way TV/video rose to the top as a significant predictor. However, the results for achievement and attitude were exactly the reverse of each other in this regard. For achievement, TV/video improved DE conditions for both synchronous and asynchronous DE, while use of the telephone favored classroom conditions in synchronous DE. For attitudes, TV/video favored the classroom and use of the telephone favored DE, both in synchronous DE settings. Generally speaking, these results appear to further implicate communication and the use of supplementary visual materials.

If one over-arching generalization is applicable here it is that sufficient opportunities for both student/instructor and student/student communication are important, possibly in the service of collaborative learning experiences such as problem-based learning. We encourage practitioners to build more of these two elements into DE courses and into classroom experiences as well. We also favor an interpretation of media features as aids to these seemingly important instructional/pedagogical aspects of course design and delivery. For DE, in particular, where media are the only means of providing collaborative and communicative experiences for students, we see pedagogy and the media that support it working in tandem and not as competing entities in the course developer/instructors set of tools. So, while we have attempted to separate pedagogy from media to assess their relative importance, it is the total package in DE that must ultimately come together to foster student learning and satisfaction.

**General considerations.** Researchers, educators and the business community have all commented recently on the future of education and the goals of schooling. These comments focus on the importance of encouraging learners to have a lifelong commitment to learning, to be responsible for their own learning, to have effective interpersonal and communication skills, to be aware of technology as a tool for learning and to be effective problem solvers with skills transferable to varied contexts. These comments also recognize that learners who have genuine learning goals are likely to remain personally committed to their achievement goals, use complex cognitive skills and draw upon the active support of the learning community to enhance their personal skills. These concerns apply with equal if not greater force to learning at a distance where the challenges of isolation may exacerbate them.

The results of this meta-analysis provide general support for the claim that effective DE depends on the provision of pedagogical excellence. How is this achieved in a DE environment? Particular predictors of pedagogical importance included problem-based learning and interactivity, either face-to-face or through mediation, with instructors and other students. Can we make a more general case? We speculate that the keys to pedagogical effectiveness in DE center on the appropriate and strategic use of interactivity among learners and with the learning material leading to learner engagement, deep processing and understanding. By what means might interactivity occur?

First, interactivity among learners occurs when technology is used as a communication device and learners are provided with appropriate collaborative activities and strategies for learning together. Here we distinguish between “surface” interaction among learners, where superficial learning is promoted through efficient communication (e.g., seeking only the correct answer), and “deep” interaction among learners, where complex learning is promoted through effective
communication (e.g., seeking an explanation). The teacher plays roles here by participating in establishing, maintaining and guiding interactive communication.

Second, the design of interactivity around learning materials might focus on notions derived from cognitive psychology including socio-cognitive and constructivist principles of learning such as those summarized by the American Psychological Association (1995, 1997). Additionally, learning materials and tasks must engage the learner in ways that promote meaningfulness, understanding and transfer. Clarity, expressiveness and feedback may help to insure learner engagement and interactivity; multimedia-learning materials may do likewise when linked to authentic learning activities.

**Considerations for Policymakers**

One possible implication is that DE needs to exploit media in ways that take advantage of its power; not just DE as an electronic copy of paper-based material. This may explain why the effect sizes are so small in the current meta-analysis. That is, there is a widespread weakness in the tools of DE. Where are the cognitive tools that encourage deeper, active learning—the ones that Kozma and Cobb predicted would transform learning experience? These need further development and more appropriate deployment. A contrasting view, supported by the size of effects encountered in this quantitative review, is that DE effectiveness is most directly affected by pedagogical excellence rather than media sophistication or flexibility.

The first alternative is a longstanding speculation that likely may not be verified until the next generation of DE is widely available and appropriately used. The second alternative requires that policymakers devote energies to ensuring that excellence and effectiveness take precedence over cost efficiency.

**Considerations for Future DE Research**

What does this analysis suggest about future DE research directions? The answer to this question depends, to some extent, upon whether we accept the premise of Clark and others that media comparison studies (and DE comparison studies, by extension) answer few useful questions, or the premise of Smith and Dillon (1999), that there is still a place for comparative studies, performed under certain conditions. It is probably true that, once established as a “legitimate alternative to classroom instruction,” the need for comparative DE studies diminishes. After all, even in the world of folklore, the comparison between a steam-driven device and the brawn of John Henry was performed only once, to the demise of John. But it is also true that before we forge ahead into an indeterminate future, possibly embracing untested fads and following false leads, while at the same time dismantling the infrastructure of the past, we should reflect upon why we are going there and what we risk if we are wrong. And if there is a practical way of translating what we know about “best practices in the classroom” to “best practices in cyberspace,” then a case for continued research in both venues, simultaneously, might be made.

So what can we learn from classroom instruction that can be translated into effective DE practices? One of the few significant findings that emerged from the TV studies of the 50s and 60s was that planning and design pay off—it was not the medium that mattered so much as what came before the TV cameras were turned on. Similarly, in this millennium, we might ask if there are aspects of design, relating to either medium or method, that are optimal in either or both instructional contexts. In collecting these studies we found few factorial designs suggesting that the bulk of the studies asked questions in the form of, “Is it this or that?” Comparisons such as this are the stock-in-trade of meta-analysis, but once the basic question is answered, more or less, we should begin to move towards answering more subtle and sophisticated questions. More
complex designs might enable us to address questions such as, “What does it depend on or what moderates between this and that?” Simply knowing that something works or doesn’t work without knowing why strands us in a quagmire of uncertainty allowing the “gimmick of the week” to become king. It is the examination of the details of research studies that can tell us the “why.”

So if comparison studies do continue—and we suspect that will—can we envisage an optimal comparative study? In the best of all Campbell and Stanley (1963) worlds, an experiment that intends to establish cause eliminates all rival hypotheses and varies only one aspect of the design—the treatment. So, here that means eliminating all potential confounds—selection, history, materials, etc.—except distance, the one feature that distinguishes distance education from face-to-face instruction. The problem is that, even if exactly the same media are used in both the DE and the classroom conditions, they are used for fundamentally different purposes, in DE to bridge the distance gap (e.g., online collaborative learning instead of face-to-face collaboration) and in the classroom as a supplement to face-to-face instruction. So, almost every time we examine the problem of media/method confounds and other sources of inequality between treatments, we have already identified a fundamental stumbling block to deriving any more useful information from comparative studies. This does not mean, of course, that imperfectly designed but perfectly described studies (i.e., descriptions of the details of treatments and methodology) are not useful in the hands of a meta-analyst, but will we learn any more than we already know by continuing to pursue comparative research. We suspect not.

So, what is the alternative? In the realm of synchronous DE a useful set of studies might involve two classroom/DE dyads, run simultaneously, with one of a host of instructional features being varied across the treatments. In a study of this sort, media are used for the same purpose in both conditions, and so distance is not the variable under study. In asynchronous DE we envisage similar direct comparisons between equivalent DE treatments. Bernard and Naidu (1992) performed a study of this sort that compared different conditions of concept mapping and questioning among roughly equivalent DE groups. Studies of this sort could even examine different types of media or media used for different purposes without succumbing to the fatal flaw that is inherent in DE/classroom-comparative research.

Some other directions for future research include:

- Developing a theoretical framework for the design and analysis of DE. Adapting the learner-centered principles of the American Psychological Association (1995,1997; see also Lambert & McCombs, 1998) may be a starting point for exploring the cognitive and motivational processes involved in learning at a distance.

- Exploring more fully student motivational dispositions in DE, including task choice, persistence, mental effort, efficacy and perceived task value. Interest/satisfaction may not indicate success, but the opposite, since students may spend less effort learning especially when they choose between DE and regular courses for convenience purposes (i.e., happy to have choice and satisfied, but because they may wish to make less of an effort to learn they are merely conveniencing themselves).

- Examining new aspects of pedagogical effectiveness and efficiency, including faculty development and teaching time, student access and learning time and cost effectiveness (e.g., cost per student). Establishing desirable skill-sets for instructors of synchronous and asynchronous DE settings might be a place to start. Examining different methods for developing these skill-sets might extend from this examination.
• Studying levels of learning (e.g., simple knowledge or comprehension vs. higher-order thinking). Examining various instructional strategies for achieving these outcomes, such as PBL and collaborative online learning, could represent a very productive line of inquiry.

• Examining inclusivity and accessibility for home learners, rural and remote learners and learners with various disabilities. Here in particular the appropriate comparison may be with “no instruction” rather than “traditional” classroom instruction.

• Using more rigorous and complete research methodologies including more detailed descriptions of control conditions in terms of both pedagogical features as well as media characteristics.

There is one thing that is certain. The demand for research will always lag behind the supply of research and for this very reason it is important to apportion our collective research resources judiciously. It may just be that at this point in our evolution, and with so many pressing issues to examine as Internet applications of DE proliferate, continuing to compare DE with the classroom is wasted time and effort.

Conclusion
This meta-analysis represents a rigorously applied examination of the comparative literature of DE with regard to the variety of conditions of study features and outcomes that are publicly available. We found evidence, in an overall sense, that classroom instruction and DE are comparable, as have some others. However, the wide variability present in all measures precludes any firm declarations of this sort. We confirm the prevailing view that, in general, DE research is of low quality, particularly in terms of internal validity. We found a dearth of information in the literature; a more replete literature could have led to stronger conclusions and recommendations for practice and policymaking. Beyond that, we have also contributed the following: 1) a view of the differences that exist in all measures between synchronous and asynchronous DE; 2) a view of the relationship between pedagogy and media that appears as a focus for debate whenever a new learning orientation (e.g., constructivism) or medium of instruction (e.g., computer-mediated communication) appears on the educational horizon; 3) an assessment of the relative strength and effect of methodological quality on the assessment of other contributing factors; 4) a glimpse of the relatively few individual study features that predict learning and attitude outcomes; and 5) a view of the heterogeneity in findings that hampered our attempts to form homogeneous subsets of study features that could have helped to establish what makes DE better or worse than classroom instruction.
References

(References with stars are studies in the meta-analysis.)


*Hilgenberg, C., & Tolone, W. (2001). Student perceptions of satisfaction and opportunities for critical thinking in distance education by interactive video. In M. G. Moore, & J. T. Savrock, (Eds.), Distance education in the health sciences University Park, PA: The American Center for the Study of Distance Education.


Lou, Y. (in press). Learning to solve complex problems through online between-group collaboration. *Distance Education.*


*Sorensen, C. K. (1996).* Students near and far: Differences in perceptions of community college students taking interactive television classes at origination and remote sites. (ERIC Document Reproduction Service No. ED 393 509)


Appendix

Coded variables and study features in the DE meta-analysis codebook

Section A: Identification of Studies

1. Study Number (Name: “Study”);
2. Finding Number (Name: “Finding”);
3. Author Name (Name: “Author”);
4. Year of Publication (Name: “Yr”);

Section B: Outcome Features

1. Outcome Type (Name: “Outcome”):
   1. Achievement
   2. Retention
   3. Attitude towards course
   4. Attitude towards the technology
   5. Attitude towards the subject matter
   6. Attitude towards the instructor
   7. Other attitudes
4. Whose Outcome (Name: “Whose”):
   1. Group
   2. Individual
   3. Teacher
5. Number of Control Conditions (Name: “Ctrol”):
   1. One control, one DE
   2. One control, more than one DE
   3. One DE, more than one control
   4. More than one DE and more than one control

Section C: Methodological Features

1. Type of Publication (Name: “Typpub”):
   1. Journal article
   2. Book chapter
   3. Report
   4. Dissertation
2. Outcome Measure (Name: “Measure”):
   1. Standardized test
   2. Researcher-made test
   3. Teacher-made test
   4. Teacher/researcher-made test
3. Effect Size (Name: “Esest”):
   1. Calculated
   2. Estimated from probability levels
4. Treatment Duration (Name: “Durat”):
   1. Less than one semester
   2. One semester
   3. More than one semester
5. Treatment Proximity (Name: “Prox”):
1. Same time period
2. Different time period

6. Instructor equivalence (Name: “Inseq”):
   1. Same instructor
   2. Different instructor

7. Student equivalence (Name: “Stueq”):
   1. Random assignment
   2. Statistical control

8. Equivalent time on task (Name: “Timeeq”):

9. Material Equivalence (Name: “Mateq”):
   1. Same curriculum materials
   2. Different curriculum materials

10. Learner ability (Name: “Abilit”):

11. Attrition Rates (Name: “Attr”):

12. Average Class Size (Name: “Size”):
   1. DE larger than control
   2. DE equal to control
   3. DE smaller than control

13. Gender (Name: “Gender”):

Section D: Course Design and Pedagogical Features

1. Simultaneous Delivery (Name: “Simul”):
   1. Simultaneous delivery
   2. Not simultaneous

2. Systematic ID “Instructional Design” (Name: “Id”):

3. DE condition: Advance Information (Name: “Advinf”):
   1. Information received prior to commencement of the course
   2. Information received at the first course
   3. No information received

4. Opportunity for F2F/instructor (Name: “f2ft”):
   1. Yes. Opportunity to meet the instructor during instruction
   2. No opportunity to meet the instructor
   3. Yes. Opportunity to meet the instructor prior to, or at the commencement of, instruction only (example: orientation session)

5. Opportunity for F2F contact/peers (Name: “f2fp”):
   1. Yes Opportunity to meet peers during instruction
   2. No opportunity to meet peers
   3. Opportunity to meet peers at or prior to the commencement of instruction

6. Provision for Synchronous technically-mediated Communication/teacher (Name: “Syncte”):
   1. Opportunity for synchronous communication
   2. No opportunity for synchronous communication

7. Provision for synchronous technically-mediated communication/students (Name: “Synper”):
   1. Opportunity for synchronous communication
   2. No opportunity for synchronous communication

8. Teacher/Student Contact Encouraged (Name: “Tstd”):

9. Student/Student Contact Encouraged (Name: “Ss”):
10. Problem-based Learning (Name: “Pbl”):*

Section E: Institutional Features

1. Institutional Support for Instructor (Name: “Insupp”):*
2. Technical Support for Students (Name: “Tcsupp”):*

Section F: Media Features

1. Use of 2-way audio conferencing (Name: “Ac”):*
2. Use of 2-way video conferencing (Name: “Vc”):*
3. Use of CMC or interactive computer classroom (Name: “Cmc”):*
4. Use of e-mail (Name: “Email”):*
5. Use of 1-way broadcast TV or videotape or audiotape (Name: “Tvvid”):*
6. Use of web-based course materials (Name: “Web”):*
7. Use of telephone (Name: “Tele”):*
8. Use of computer-based tutorials/simulations (Name: “Cbi”):*

Section G: Demographics

1. Cost of Course Delivery (Name: “Cost”): *
2. Purpose for Offering DE (Name: “Purpos”):
   1. Flexibility of schedule or travel
   2. Preferred media approach
   3. Access to expertise (teacher/program)
   4. Special needs students
   5. Efficient delivery or cost savings
   6. Multiple reasons. Specify:
3. Instructor Experience with DE (Name: “Inde”):
   1. Yes
   2. No
4. Instructor Experience with technologies used (Name: “Intech”):
   1. Yes
   2. No
5. Students’ Experience with DE (Name: “Stude”):
   1. Yes
   2. No
6. Students’ Experience with technologies used (Name: “Stutech”):
   1. Yes
   2. No
7. Types of Control Learners (Name: “Lrtypc”):
   1. Grade School (K-12)
   2. Undergraduate
   3. Graduate
   4. Military
   5. Industry/business
   6. Professionals (e.g., doctors)
8. Types of DE Learners (Name: “Lrtypd”):
   1. Grade and high school (K-12)
   2. Undergraduate
   3. Graduate
4. Military
5. Industry/business
6. Professionals (e.g., doctors)

9. Setting (Name: “Setting”):
   1. DE urban and control rural
   2. DE urban and control urban
   3. DE reported/control not reported
   4. DE rural and control urban
   5. DE rural and control rural
   6. Control reported/DE not reported

10. Subject Matter (Name: “Subjec”):
    1. Math (including stats and algebra)
    2. Languages (includes language arts and second languages)
    3. Science (including biology, sociology, psychology & philosophy)
    4. History
    5. Geography
    6. Computer science (information technology)
    7. Computer applications
    8. Education
    9. Medicine or nursing (histology)
    10. Military training
    11. Business
    12. Engineering
    13. Other (specify)

11. Average age (Name: “Age”):
    1. Real difference in age means with the corresponding sign

* These items were coded using the following scheme:
  1. DE more than control group
  2. DE reported/control group not reported
  3. DE equal to control group
  4. Control reported/DE not reported
  5. DE less than control group

Missing (no information on DE or control reported)