Student Achievement in Online Distance Education Compared to Face-to-Face Education

Namsook Jahng, PhD. Student [namsook@interchange.ubc.ca]

Don Krug, Associate Professor [don.krug@ubc.ca]

Zuochen Zhang, PhD. Student [zuochenz@yahoo.com]

Faculty of Education, University of British Columbia
2125 Main Mall, Vancouver, BC, V6T 1Z4

Abstract

A meta-analysis was conducted to synthesize existing research published between 1995 and 2004 comparing student achievement in online distance education (ODE) and face-to-face education (F2FE) at the post-secondary level. The purpose of this study was to investigate how the development of technology contributed to student achievement in ODE within the last ten years. The result of comparing overall weighted mean effect size of student achievement showed no significant difference between the two settings (d = .023, k = 20, N = 1617, p = 0.640). However, the student achievement comparison revealed an interesting result when the primary studies were categorized by whether the experimental study conducted a pre-test or not. In the pre-tested group of studies, student achievement in ODE was significantly higher than F2FE (d = 0.211, k = 9, N = 631, $p < 0.05$) even though there was no difference for prior knowledge between ODE and F2FE (d = 0.0813, k=9, N=631, $p>0.05$). On the other hand, student achievement from the no pre-test group of studies resulted in no significant difference between the two settings (d = -0.106, k = 11, N = 986, $p > 0.05$). Discussion and suggestion for further studies are provided focusing on methodological weakness of primary studies and differences of teaching and learning in ODE and F2FE.

Keywords

Meta-analysis, Effect size, Online Distance Education (ODE), Face-to-Face education (F2FE), Student achievement, Pre-test

Introduction

The increasing sophistication and affordability of technology has fostered the rapid growth of distance education (DE) at the post-secondary level. Online distance education (ODE) began to spread from 1995 along with the increased accessibility to the Internet (Bates & Pool, 2003). Through its historic expansion, the current and most prevalent delivery format of DE is ODE (National Education Association, 2000). While most of post-secondary education institutions are interested in developing and delivering ODE courses or programs, the awareness of the quality of learning is getting more and more important.

Some researchers believe advancement of technology influences a better quality of learning, others insist there is no benefit from different kinds of media. Regarding the quality of learning through ODE, much of the research has concluded that learning in ODE is as good as the learning in face-to-face education (F2FE) (Hong, 2002; Kleinman, & Entin, 2002; Rovai, 2002). On the other hand, contradictory findings from individual studies have reported on student achievement between the two settings. Even the quantitative and qualitative reviews of synthesizing the individual experimental studies have reported inconsistent results. Thus, more comprehensive and intensive analysis is required to examine the difference in student achievement in ODE compared to F2FE.

No Significant Difference between Distance Education and Face-to-Face Education

In the literature there exist various results from individual comparative studies on whether student achievement in DE is better or worse than in face-to-face education (F2FE). A good many reviews of the literature on the distance education effectiveness have concluded that distance education courses as effective as face-to-face courses (Allen, Bourihis, Burrell, & Mabry, 2002; Bernard et al., 2004; Cavanaugh, 2001; Machtmes & Asher, 2000; Mayzer & Dejong, 2003; Murphy, 2000; Phipps & Merisotis, 1999; Ramage, 2002; Russell, 1999; Schulman & Sims, 1999; Zhao & Tan, 2004). These studies reported that student achievement in DE can be as good as that in F2FE. Furthermore, some recent studies reported higher achievement in distance education than F2FE (e.g. Shachar and Neumann 2003; Bernard et al., 2004). Some studies also assert a more positive trend of student achievement in distance education settings than in F2FE settings as time goes by (Machtmes & Asher, 2000; Zhao et al., 2004).

Russell's (1999) research was the most frequently cited and also most criticized study in the literature on distance education. He collected 355 studies ranging from 1928 to 1998 and covering all academic levels and delivery media types from correspondence to Web-based courses. He compiled the extensive studies and listed the findings of the results, concluding that 90 percent of the primary studies reported *no significant differences* between distance education and F2FE in terms of student achievement. Russell (1999) asserted that no matter what kinds of media were involved distance education was as effective as F2FE.

Phipps and Merisotis (1999) sampled about 40 studies that were articles, essays, and other writings and they reported that, "Most of these studies conclude that regardless of the technology used, distance learning courses compare favourably with classroom-based instruction and enjoy high student satisfaction" (Phipps & Merisotis, 1999, p. 13).

Higher Student Achievement in Distance Education than Face-to-Face Education

Shachar and Neumann's (2003) research compared students' final grades in distance education with those in traditional classes. Their meta-analysis of 86 studies between 1990 and 2002 resulted in an overall effect size of 0.37. They concluded that the overall effect size was a significant difference adopting their alternative hypothesis, "The final academic performance grades of students enrolled in distance education programs are higher than those enrolled in traditional F2F programs" (ibid, p. 5).

Bernard et al. (2004) reviewed literature of empirical studies between 1985 and 2002 focusing on achievement, student attitudes, and retention rates. They investigated the effectiveness of distance education compared with its classroom-based counterparts, analyzing 688 effect sizes from 232 studies across all academic levels, media types, instructional methods and outcome measures. Accordingly, they found that there was a small but significant effect favouring distance education conditions ($g = 0.0128$, $k = 318$) on overall achievement outcomes.

It is also noteworthy that the two meta-analysis studies by Machtmes and Asher (2000) and Zhao et al. (2004) established that there was a significant relationship between the publication year and a positive effect in distance education. Machtmes and Asher (2000) examined literature on experimental studies on *telecourses* in adult and higher education. They reviewed 19 comparative studies ranging from 1943 to 1997. They coded the studies according to the decade that the study was conducted and found that each successive decade had a larger impact on learner achievement. They showed that "The effect sizes increased and moved toward positive values as the decades progressed: 1960s = -0.09; 1970s = -0.20; 1980s = +0.04; and 1990s = +0.23" (Machtmes & Asher, 2000, p. 40). The authors argued that these results might be caused by technological advances in the media that allowed for greater interaction and various presentation styles, concluding that technology was clearly the most significant difference between these decades: not only were effect sizes more positive in the latter two decades, but became greater moving from the 1980s to 1990s. Machtmes and Asher (2000) attributed this change of effectiveness to improvements of technology such as CD-ROM and Web-based instruction.

Zhao et al. (2004) sampled 51 articles from peer-reviewed journals published between 1966 and 2002. The overall weighted mean effect size was +0.10, with a 95% confidence interval of [-.01 .22] ($z = 1.76$, $p > .05$, $SD = 0.06$). They confirmed previous research findings (e.g. Clark, 1983; Russell, 1999; Phipps & Merisotis, 1999; Machtmes & Asher, 2000) that reported "no significant difference between distance education and F2FE" (Zhao et al., 2004, p. 28). When they categorized the primary studies by publication years, however, they found that "significantly positive effectiveness" ($d = 0.20$, $k = 77$) in distance education from the articles published after 1998 while there was no significant difference between the two settings found in the publications prior 1998 ($d = -0.10$, $k = 20$). They interpreted the finding as a possible influence from the significant changes in technology that were employed in distance education around the mid-1990s such as learning material presentation and communication through World Wide Web and the Internet.

Debate on the Influence from Various Media on Student Achievement

Richard Clark and Robert Kozma represent the opposite ends of a debate regarding the influence of media on the effectiveness of education. Clark (1983) concluded that "Consistent evidence is found for the generalization that there are no learning benefits to be gained from employing any specific medium to deliver instruction" (p. 445). He argued that media are mere vehicles that do not influence learning. Clark's (1983, 1994) assertion has been supported and confirmed by many researchers such as Bernard et al. (2004), Phipps and Merisotis (1999), Rovai (2002), and Zhao et al. (2004).

A decade later, Kozma (1994) revisited Clark's claim and questioned "Will media influence learning?" In terms of "media's capabilities" (p. 17), he argued that considerable changes had been made since the 1980s. Kozma (1994) pointed out the samples Clark (1983) had reviewed were too old to assess the effectiveness of recent media because most of them were published before the 1980s. Anticipating the advent of more interactive media in the near future, Kozma (1994) expected more abrupt changes. He claimed that "If there is no relationship between media and learning it may be because we have not yet made one" (p. 7). Refuting Clark's delivery truck metaphor, that media simply conveys content, Kozma (1994) argued on the basis of constructivism - that learning is not a receptive response to instruction's delivery. Shuell (1988) stated that "learning is an active, constructive, cognitive and social process by which the learner manages available cognitive, physical, and social resources to create new knowledge by interacting with information in the environment and integrating it with information already stored in memory" (cited in Kozma, 1994, p. 8).

With Kozma's view of abrupt changes in more interactive media since mid-1990s, the Web media of today can be substantially different from older media. In this respect, Kozma's assertion on the role of media should be re-evaluated with comprehensive review based on comparative experimental studies published since 1995, the time when the Web technologies were introduced in distance education.

Methodological Design Flaws in Comparative Studies Included in Review Studies

Phipps and Merisotis (1999) identified critical methodological flaws in primary studies that influenced the overall quality of control of extraneous variables, sampling, validity and reliability of the experiments. Benard et al. (2004) asserted that the wide distribution of effect sizes precluded any firm declarations of the effectiveness of distance education. They insisted that the tremendous variability of effect sizes were from a methodological weakness in distance education research. Although the quality of primary studies included meta-analysis studies (e.g. Cavanaugh, 2001; Matchmes & Asher, 2000) was ostensibly higher than those complied in Russell's (1999) review because of careful screening procedures in sampling studies, the low quality of methodological design in primary studies was still blamed for the significant heterogeneity of the effect sizes (Zhao et al., 2004).

Methodology

Meta-analysis

Meta-analysis is a quantitative method to integrate and summarize the findings of individual studies (Glass, 1976). He stated "Meta-analysis refers to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976, p. 3). Its basic purpose is to apply the same methodological rigor to a literature review that is used in primary experimental research through coding and interpreting collected studies (Cooper & Hedges, 1994).

With an increase in the number of quantitative comparative studies of the effectiveness of distance education, since the 1980s, there have been efforts to synthesize information using meta-analysis.

Meta-analysis is well-suited for analyzing the influence from moderator variables: It clearly indicates how methodological variations influence the strength of an effect, that is one of the strong points of meta-analysis (Cooper & Hedges, 1994).

Data Sources

Identifying relevant studies is one of the most important procedures in meta-analysis since missing data and publication bias might endanger the validity of the study (Hedges & Olkin, 1985). Publication bias refers to the greater likelihood, on average, of a study with positive significance to be published (Olson et al., 2002). In order to avoid this bias, researchers have included unpublished studies, as well as published articles, when conducting meta-analysis.

On the other side of the publication bias, however, there is a problem concerning the quality of sampled studies. The low quality of research design in comparative studies, those with a lot of methodological flaws, have been invariably pointed out as the most serious obstacle for adopting review findings (Benard et al., 2004; Clark, 1983; Joy & Garcia, 2000; Phipps & Merisotis, 1999).

In relation to these two problems (i.e. publication bias and low quality of research design), we decided to select only studies published in journals. we agreed with Zhao et al. (2004) that "only including journal articles may result in publication bias but we believed that the risk was minimal as there had not been a dominant paradigm for distance education over the years to cause a certain bias against or for positive, negative, or non-significant findings" (p. 10).

However, Zhao et al.'s (2004) research method also risked missing relevant studies because only the Educational Resources Information Center (ERIC) was searched for data. In order to avoid this problem, we conducted a comprehensive search through three major sources for data searching:

1. electronic databases,
2. major journals in distance education, and
3. reference lists of previous studies.

Initially, we searched three electronic databases through the University of British Columbia's library website (ERIC, ACM digital library, and PsycInfo) employing the following keywords: distance education and effectiveness, distance education and achievement/learning/outcome. In addition, online education or Web-based education was substituted for distance education.

Manual searches were also performed in three major journals in distance education: American Journal of Distance Education (AJDE), Journal of Distance Education (JDE), and Journal of Asynchronous Learning Network (JALN). The AJDE provides only titles and abstracts of articles online, while the JDE provides full text for articles from back issues. The JALN is a free online journal.

The third source was the reference list of previous meta-analysis studies. In particular, three recent meta-analyses (Bernard et al., 2004; Ungerleider & Burns, 2003; Zhao et al., 2004) were important sources. The citation information of all the retrieved studies was put into the software program EndNote (7.0) to build a database.

Inclusion / Exclusion Criteria for this Meta-analysis

In this meta-analysis, studies had to meet the following criteria to be included:

1. Delivery was web-based (excluded interactive television course, video/audio delivery distance course, correspondence course, etc.);
2. Learners were in post-secondary levels of higher education;
3. Articles were published between 1995 and 2004;
4. Studies were publicly accessible (excluded unpublished dissertations, reports or documents);
5. Experimental or quasi-experimental studies included enough statistical information for computing effect sizes. (excluded case studies, position papers);
6. There were comparative outcomes between the control (traditional face-to-face instruction) and experimental (online distance education) groups. Experimental results only from distance education courses that did not have comparison group were excluded; and,
7. Studies were published in English.

Data Selection Procedure

The first step of data selection was using keywords to search electronic databases. Keywords were identified from a review of previous studies. The extensive search resulted in thousands of articles, reports, surveys, opinion papers etc. Studies were entered into EndNote (7.0) to build a master candidate database.

Next, we screened the studies by title and publication year. This process reduced the number of possible candidates to about 500 studies. These data were transferred again to EndNote (7.0) to build a second database. We filtered the publications in the second database by reading the abstracts. This process resulted in 121 manuscripts for inclusion in the third database. We also conducted a manual search of three major journals that yielded about 30 studies. By examining the reference lists of previous studies (e.g. Bernard et al., 2004; Ungerleider & Burns, 2003; Zhao et al., 2004), we identified an additional 78 articles. After deleting duplicates, 153 manuscripts were included in the fourth database. We collected the manuscripts and read them carefully to select the studies that met the inclusion and exclusion criteria. Twenty studies from nineteen articles were selected that met all the inclusion criteria of this meta-analysis.

Coded Variables

The articles were coded using four categories, i.e., citation, research design, ODE course design, and student achievement (see Table 1). We used final test scores for student achievement. Information resulting from the experiments was coded to calculate effect sizes of student achievement (e.g. sample sizes, mean scores, ratios, t-values, p-values, etc.).

In order to analyze the factors that are related to student achievement, we coded course information (course name, subject area, and level of course), student characteristics (average ages, gender ratio, and level of technology skill), and instructor information (level of technology skill, ODE teaching experience).

Methodological features were also coded to examine the validity of the experiment related to the controlling of extraneous variables. This category was based on the assumption that *more controlled experiments would result in smaller variances of effect sizes than less controlled experiments*. For example, results of final test scores should show a smaller variance when the courses designed with the same teaching material, the same test methods, and the same instructors for experiments than with other experiments using different materials, test methods and instructors. On the basis of this methodological hypothesis, a pre-test variable was used to examine the quality of experimental design.

Table 1. Coding Categories and Variables

Category	Variables		Description
Citation	Author name		Name of the first author
	Year		Published year of an article
	Publication source		Journal name
Research design	Course feature	Course name	Name of study course
		Course subject area	Subject fields classified
		Course level	Graduate or undergraduate course
	Student feature	Student age	Average ages of students
		Student gender	Gender ratio
		Student technology	Level of technology skill
	Instructor feature	Instructor technology	Level of technology skill
		Instructor experience	Distance education course teaching experiences
	Methodological feature	Type of experiment	quasi-experiment or not
		Sampling method	Random or self-selection
		Sample size	Number of students in ODE/F2F
		pre-test	pre-test scores
		Same material	Same material for the both formats
		Same instructor	Same instructor taught or not
		Same test	Test with same quizzes or exams
ODE course design	Delivery application		Kind of courseware (WebCT, etc.)
	Types of communication		Synchronous or Asynchronous
	Frequency of class		Class times Per week
Outcome	Student achievement		Final test scores

Effect Size Calculation

An effect size is a degree of standardized mean difference between the experimental group (ODE) and the control group (F2FE) (Bernard et al., 2004). Effect sizes for each outcome variable were identified from each individual study and then compiled.

The population effect size δ was defined as:

$$\delta = (\mu_e - \mu_c) / \sigma,$$

where μ_e and μ_c are the population mean of the experimental group and that of the control group, respectively, and *s* is the standard deviation (SD) of the population (Glass, 1976). The population effect size was estimated by Cohen's *g*, defined as:

$$g = (M_e - M_c) / S_{pooled},$$

where M_e (M_c) was the sample mean of the experimental group (the control group) and S_{pooled} was the pooled SD of the sample deviations (Cooper & Hedges, 1994). Cohen's *g* was adjusted with a correction term C_m

$$C_m = 1 - 3 / (4n_e + 4n_c - 9),$$

where n_e (n_c) was the sample size of the experimental group (the control group) because *g* was a biased estimator of δ (Hedges & Olkin, 1985). Thus, we used the Hedges' *d* as the unbiased estimator of the population effect size as

$$d = C_m \times g.$$

After calculating each effect size (*d*) from the individual study, we combined effect sizes to report the central tendency of the data. The simplest method of combining *d*s is to average them. However, averaging after weighting effect sizes is more appropriate method, because different studies estimate the true effect size with varying degree of precision. Thus, we combined *d*s using a weighted average method (i.e. the inverse variance-weighted method):

$$\bar{d} = \sum w_i d_i / \sum w_i ,$$

where w_i = 1/ variance of d_i. Setting w_i to be the inverse variance of a study is more frequently used than setting w_i to be the sample size of a study, because the degree of precision of a study can be represented better with the variance of effect sizes (Cooper & Hedges, 1994).

Data Analysis

The homogeneity test was important since the results from all the individual experiments were synthesized to produce a result based on the assumption that they were from the same population. In order to test the homogeneity of a sample of effect sizes, statistics Q was calculated and checked whether the studies shared a common effect sizes with testing Q by a Chi-square (χ^2) distribution of k-1 degree of freedom, where k was the number of effect sizes (Hedges & Olkin, 1985). A five percent significance level was used for the homogeneity test. When the homogeneity statistics Q was rejected, the mean effect size had to be reanalyzed after eliminating a non-homogeneous study of the farthest effect size from the previous synthesized effect size. The homogeneity tests were repeated until there were no heterogeneous experiments in the sample.

Zero mean effect size indicated that there was no difference between the experimental and control groups (Machtmes & Asher, 2000). If the study produced a negative (-) effect size, this meant that the control group (traditional face-to-face instruction in this study) was more effective than the experimental group (distance education in this study) - A positive (+) effect size meant the opposite result. A five percent significance level was used for testing significant difference of an effect size from zero and to calculate the lower and upper limits by 95 percent confidence interval. Although the interpretation of an effect size was heavily dependent on the area that meta-analysis was applied, as a rule of thumb, we considered a major difference when any effect size was greater than ±0.50 and a considerable difference with a ±0.20 to ±0.50 effect size (Hedges & Olkin, 1985).

All coded information was stored in a Microsoft Excel program. Calculating effect sizes for individual studies was conducted using *Comprehensive Meta Analysis Version 1.0.23* (software for meta-analytic review of literature). Later the data in these two software programs were imported into *SPSS* (Statistical Package for the Social Sciences) *Version 8.0* for further analyses.

Results

Data Characteristics

As Table 2 shows, this study examined student achievement in ODE compared to F2FE with a time frame from 1995 until 2004. The selection procedures identified twenty studies (k=20, N=1617) from nineteen articles that had enough information to calculate effect sizes of student achievement. Waschull (2001) received two different study IDs, 'Waschull01s1' and 'Waschull01s2', because of conducting two different experiments that contained two results in one article. The final test score was used for effect size calculation, instead of a final grade, because a final grade usually integrated all activities during a course including attendance and participation as well as test scores. Studies that did not report a final test/exam score, but presented test scores during the course were included after we calculated the average of this data (e.g., Leasure, Davis, & Thievon, 2000).

In terms of calculating effect sizes in meta-analysis, there were several different formulas that we could have used depending on the statistical features of the studies presented such as mean, SD, t-value and p-value. For the studies that reported more features than those needed to calculate an effect size, we gave priority to using mean and standard deviation (SD) and then p-value (e.g., Navarro & Shoemaker, 1999). However, Wang and Newlin's (2000) data were not reasonable for normal distribution when mean and SD were used for calculating an effect size, confirming a similar claim by Ungerleider and Burn (2003). But the effect size was appropriate for the normal distribution when the t-value was used to calculate it. Thus, their study was included in achievement analysis.

Nine (k=9, N=631) out of the twenty studies for achievement analysis presented pre-test scores. Yes-pre-test studies were categorized as an important group to compare with the result of achievement with No-pre-test studies. This category was interpreted as a factor in relation to the arguments of research design flaws mentioned in almost every review of literature during the last few decades. Clark (1983) pointed to the problem of low quality of research design for experimental studies, arguing for control of extraneous variables that influence the validity of comparative studies. We believe that the control prior knowledge is especially needed for comparing achievement, when measuring knowledge gained through a course.

Table 2. Summary of Data Characteristics

Variables			Number of Studies (k) (students=N)
Student Achievement k=20 (N=1617)	Pre-test	Yes	k=9 (N=631)
		No	k=11 (N=986)
	Course level	Undergraduate	k=15 (N=1438)

		Graduate	k=5 (N=179)
	Course subject area	Computer Related	k=3 (N=414)
		Business	k=5 (N=289)
		Math & Science	k=5 (N=479)
		Social science & others	k=7 (N=435)
	Comparison method	Mean, standard deviation (SD)	k=17 (N=1376)
		t-value, p-value	k=3 (N=241)
	Publication year	2003~2000	k=15
		1999~1997	k=5

For the categories of course level, more studies were located from the undergraduate level (k=15, N=1438) than the graduate level (k=5, N=179) for student achievement analysis. Two articles (Schoenfeld-Tacher, McConnell, & Graham, 2001; Waschull, 2001) did not clearly indicate a course level and we estimated the level based on the course name, course characteristics, and the institutions name. Course subjects were grouped into four areas (i.e. computer related subjects, business, math & science, and social science & others).

The time frame for the inclusion criteria was post-1995 until 2004 consisting of ODE. All the studies were found in journals after 1997, we assumed the two year gap reflected the delay in publishing research.

Overall Student Achievement

The overall combined effect size for student achievement was d = 0.0234 (k=20, N=1617) with a 95% confidence interval of [-0.0797, +0.1264] as shown in Table 3. This indicated that students received slightly higher final test scores in ODE, but there was no significant difference between the two settings. A closer look at the individual effect sizes from twenty studies revealed that exactly half of them (ten studies) reported positive results for distance education and the other half had negative effect sizes. Interestingly, three of six studies (Maki02, Schoenfeld01, Tucker01) reported a significant difference in favour of ODE, while the other three studies (Gainor02, Wang00, Waschull01s1), reported significant achievement in F2F settings. This result illustrated how individual comparative studies have reported the contradictory results of student achievement between distance education and F2FE. Some meta-analysis studies reported positive effect sizes for distance education (Liao, 1999; Shachar & Neumann, 2003) while most presented no significant differences between the two educational settings (Phipps & Merisotis, 1999; Russell, 1999).

We determined a 5% confidence level for the homogeneity test, which means that data were considered homogeneous when p-value was bigger than 0.05 with Chi-square test. The homogeneity test with 20 achievement effect sizes resulted in $\chi^2(19)=44.716$ (p=0.0008). The fact that the p-value was close to zero represented how heterogeneous the effect sizes were. Based on the homogeneity tests, four studies (Waschull01s1, Gainor02, Tucker01, and Schoenfeld01) had to be removed and the results were recalculated. The final weighted mean effect size was d = 0.0259 with a 95% confidence interval of [-0.0828, +0.1346] and $\chi^2(15)=24.795$ (p>0.05).

Table 3. List of Effect Sizes for All Studies

Study ID	Effect Size (d)	Standard Error (SE)	95% Confidence Interval		p-value
			Lower Limit	Upper Limit	
Waschull01s1[1]	-0.844	0.367	-1.564	-0.124	0.022*
Gainor02[2]	-0.681	0.307	-1.283	-0.079	0.027*
Wang00	-0.543	0.192	-0.919	-0.167	0.005*
Navaro99	-0.45	0.255	-0.95	0.051	0.078
Waschull01s2	-0.433	0.318	-1.057	0.191	0.174
Wegner99	-0.268	0.363	-0.979	0.442	0.459
Schaik03	-0.17	0.292	-0.743	0.402	0.56
Johnson00	-0.116	0.369	-0.838	0.607	0.754
Collins00	-0.108	0.228	-0.555	0.34	0.637
Weems02	-0.071	0.348	-0.754	0.612	0.838
Schulman99	0.013	0.205	-0.389	0.414	0.951
Ocker99	0.103	0.633	-1.137	1.344	0.871
Smeaton97	0.126	0.129	-0.127	0.378	0.33

Leasure00	0.133	0.277	-0.409	0.675	0.631
Sankaran00	0.135	0.19	-0.237	0.508	0.477
JohnsonM02	0.236	0.188	-0.132	0.605	0.209
Tesone03	0.275	0.245	-0.206	0.755	0.262
Maki02	0.367	0.147	0.079	0.655	0.012*
Schoenfeld01(4)	0.713	0.359	0.008	1.417	0.047*
Tucker01(3)	0.724	0.301	0.133	1.314	0.016*
Overall Q(19)=44.716 (p=0.0008)	0.0234	0.0526	-0.0797	0.1264	0.6566
Final Q(15)=24.795 (p=0.0528)	0.0259	0.0555	-0.0828	0.1346	0.6400

Study ID's represent the first author's last name and last two digits from the of publication year.

* presents a significant difference (p<0.05).

(1),(2),(3) and (4) indicate the order of elimination from the homogeneity test.

However, when the mean effect size was calculated from combining only the homogeneous studies, it was nearly the same as that of all studies including the heterogeneous studies: There existed a little positive effect for ODE but the difference was not significant. In brief, this result confirmed the previous findings of quantitative and qualitative reviews of literature that resulted in *no significant difference* in student achievement between the two settings.

Prior Knowledge

The effect sizes of pre-test scores from nine studies (N=631) are displayed in Table 4. All of the nine studies were published after 2000. There were two studies (Johnson00, JohnsonM02) resulted a significant difference in prior knowledge. Seven studies showed no significant difference for the results of pre-tests between ODE and F2FE. The weighted mean effect size for the pre-test scores was d = 0.0813 (p>0.05) demonstrating no significant difference in prior knowledge between ODE students and F2FE. There was also no heterogeneous study that met χ^2 (8) =15.2171 within the five percent level of significance. In other words, all the Yes-pre-test studies consistently reported no significant gap in student content knowledge before they began the courses. This was an unexpected finding because some difference was presumed, considering the changing social and cultural characteristics of students in ODE courses compared with those students in F2F settings.

Table 4. Effect Sizes for Yes-pre-test Studies

Study ID	Effect Size (d)	Standard Error (SE)	95% Confidence Interval		p-value
			Lower Limit	Upper Limit	
Johnson00	-.6559	0.333	-1.309	-0.003	0.049*
Schoenfeld01	-.3972	0.464	-1.306	0.511	0.392
Schaik03	-.3265	0.239	-0.796	0.143	0.173
Gainor02	-.2181	0.300	-0.805	0.369	0.467
Maki02	.1186	0.152	-0.180	0.417	0.436
Tesone03	.2703	0.245	-0.210	0.751	0.270
Tucker01	.3121	0.294	-0.263	0.887	0.288
JohnsonM02	.3776	0.189	0.007	0.748	0.046*
Weems02	.5330	0.354	-0.162	1.228	0.133
Final (k=9, N=631) Q(8)= 15.2171 (p=0.0551)	0.0813	0.0818	-0.0790	0.2415	0.3204

Study ID represents the first author's last name and last two digit numbers of publication year.

* presents a significant difference (p<0.05).

ODE students were identified in the literature as non-traditional adult students who were mature, more independent than traditional students, married, female, and full-time workers (Jahng & Krug, 2003, Dutton, Dutton & Perry, 2002). Some researchers asserted that students enrolled in distance education courses were more academically motivated and had a higher Cumulative Grade Point Average (CGPA) (Hong, 2002). However, the weighted mean effect size of pre-test scores presented no significant difference at least with respect to prior knowledge between student groups in the different settings.

Achievement Outcomes of Yes-pre-test and No-pre-test Studies

There was no significant difference found for the overall mean effect size of student achievement outcomes (k=20) and for the prior content knowledge measured with pre-tests (k=9) between the two educational settings. However, we found a significant difference for weighted mean effect sizes of student achievement outcomes when the studies were classified with a Yes- or No-pre-test. There were eleven No-pre-test studies and nine Yes-pre-test studies.

Results indicated a more positive achievement effect size from the Yes-pre-test group of studies than from the No-pre-test group of studies. Learners sampled in the Yes-pre-test studies had significantly higher scores in ODE, whereas there was no significant difference in the No-pre-test studies.

The effect size comparison for achievement in the Yes-pre-test group indicated that the ODE settings attained d = 0.211 (p < 0.05, k = 9, N = 631) based on a heterogeneous sample with $\chi^2(8) = 17.63$, $p < 0.05$ (see Table 5). When the mean effect size was recalculated, after excluding a non-homogeneous study (Gainor02), it was d = 0.279 (p < 0.05, k = 8, N = 586) and the homogeneity was satisfied ($\chi^2(7) = 8.57$, $p >$ 0.05). This result indicated that the student achievement in ODE was significantly higher than that of F2FE within a 95% significance level ranging from +0.112 to +0.446.

The weighted mean effect size in achievement from the No-pre-test studies resulted in d = -0.106 (p>0.05, k=11, N=986) based on a homogeneity test of $\chi^2(10)=18.31$ (p<0.05), that represented a heterogeneous sample of studies. After eliminating the non-homogeneous study (Waschull01s1), a reanalysis generated a higher effect size d = -0.080 (p>0.05, k=10, N=953) from the homogeneous set of studies ($\chi^2(9)=14.13$, p>0.05). This result showed that there was a negative mean effect size for ODE, but the difference between the two educational settings was not significant.

Because there was a large gap (d = 0.317) of overall effect sizes between the Yes-pre-test studies (d = 0.211) and the No-pre-test studies (d = -0.106), we conducted further analysis to test the degree of differences with ANOVA (see Table 6). The ANOVA analysis indicated that a significant difference existed in the achievement scores between the two compared groups (F=6.501, p<0.05).

Table 5. Mean Effect Size Comparison between Yes-pre-test Studies and No-pre-test Studies

		Effect size		95% Confidence interval		Homogeneity	
		d	SE	Lower	upper	Q-value	p-value
Yes-pre-test studies	**Overall(k=9, N=631)**	0. 211	0.082	0.049	0.372	17.63	0.024
	Final(k=8,N= 586)	0.279	0.085	0.112	0.446	8.57	0.285
No-pre-test studies	**Overall(k=11, N=986)**	-0.106	0.068	-0.240	0.028	18.31	0.050
	Final(k=10, N=953)	-0.080	0.070	-0.216	0.059	14.13	0.118

Table 6. ANOVA Test between the Yes-pre-test Studies and No-pre-test Studies

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0.621	1	0.621	6.501	0.021
Within Groups	1.530	16	.096		
Total	2.151	17			

Categorical Analysis of Student Achievement

Course Subject Areas. Zhao et al. (2004) classified content areas into nine groups. They reported a positive significance within p < 0.01 for distance education in four content areas: "business (d = 0.13), computer science (d = 0.48), medical science (d = 0.36) and other areas (d = 0.48). On the other hand, they reported no significant effects in distance education in the course areas of social science, science, skills, military, and math" (Zhao et al. 2004, 6). These two meta-analysis studies concluded that the course subject area was a moderator variable that affected student achievement in different educational settings. The authors reported that distance education seemed to be more appropriate for certain content areas than F2FE.

Table 7. List of Effect Sizes Depending on the Subject Areas

Subject area	Analysis step	Effect size		95% Confi. Int.		Homogeneity	
		d	SE	Lower limit	Upper Limit	Q-value	p-value
Computer related	Overall/Final (k=3,N=414)	0.093	0.100	-0.103	0.290	0.92	0.6297

Business (Tucker01[1])	Overall (k=5,N=289)	0.089	0.120	-0.146	0.325	9.61	0.0476
	Final (k=4,N= 242)	-0.030	0.131	-0.287	0.226	4.34	0.2274
Social science & others (Waschull01s1[2], Gainor02[3])	Overall (k=7, N=435)	-0.018	0.099	-0.212	0.176	19.14	0.0039
	Final (k=5, N=357)	0.138	0.109	-0.076	0.351	7.38	0.1171
Math & Science (Schoenfeld01[4], Wang00[5])	Overall (k=5, N=479)	-0.058	0.105	-0.263	0.149	13.48	0.0091
	Final (k=3, N=322)	0.072	0.134	-0.190	0.335	1.55	0.4603

Note: The numbers [1,2,3,4,5] represent the studies excluded from the homogeneity tests.

Table 8. ANOVA Test of Subjects Difference with Homogeneous Samples

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	0.021	3	0.0071	0.095	0.961
Within Groups	0.824	11	0.0749		
Total	0.854	14			

In our study computer areas had the strongest support for ODE with *d* = 0.093 based on a homogeneity set of three studies but the difference was not significant (p > 0.05) (see Table 7). This was followed by the business area with *d* = 0.089 (p>0.05), that combined five studies including a non-homogeneous study (Tucker01). But the effect size became negative when the homogeneity sample was excluded. The social science (*d* = -0.018) and math & science (*d* = -0.058) attained a more positive achievement in F2FE including heterogeneous studies (Waschull01s1, Gainor02, Schoenfeld01, Wang00). Reanalyzing the effect sizes after removing the non-homogeneous studies, both areas showed positive effects for ODE.

The ANOVA test also indicated that there were no significant differences for student achievement depending on different subject areas (see Table 8). In summery, all subject areas did not indicate significant difference for student achievement between ODE and F2FE. Except for computer related areas, all the subject areas reversed the signs of effect sizes after excluding non-homogeneous studies with homogeneity tests. However, this result should be cautiously interpreted because the sample sizes for each subject area were small and there existed large gaps among effect sizes showing non-homogeneous results.

Course Level. Student achievement in ODE was strongly related to course levels as shown in Table 9. While graduate courses were significantly less effective in ODE than those in F2FE (*d* = -0.386, p < 0.05), undergraduate courses in ODE showed higher achievement than in F2FE (*d* = 0.163, p<0.05) with a homogeneous study sample after excluding heterogeneous studies (i.e. Waschull01s1, Wang00).

Table 9. Weighted Mean Effect Sizes Categorized by Course Level

		Effect size		95% Confi. int.		Homogeneity	
		d	SE	lower	upper	Q-value	p-value
Undergraduate (Waschull01s1[1], Wang00[2])	**Overall (k=15,N=1438)**	0.079	0.056	-0.031	0.189	14.00	0.0019
	Final (k=13,N=1290)	0.163	0.059	0.046	0.279	12.00	0.2193
Graduate	**Overall/Final (k=5, N= 179)**	-0.386	0.152	-0.683	-0.088	4.00	0.6945

Waschull01s1[1], Wang00[2] represent the studies excluded from the homogeneity tests.

This finding confirmed the results of the previous meta-analysis studies that reported more positive effect sizes for ODE over F2FE at an undergraduate level than at a graduate level. For example, Zhao et al. (2004) reported that the effect size of achievement from an undergraduate level (*d* = 0.36, p<0.01) was much higher than that from a graduate level (*d* = 0.03). Similarly, Bernard et al.'s (2004) meta-analysis study stated that the effect size of achievement was more positive for the undergraduate level (*g* = 0.130, p>0.05) than the graduate level (*g* = 0.098, p>0.05).

Zhao et al. (2004) contended that undergraduate courses usually encompass the acquisition of content knowledge and skills while graduate courses tend to deal with higher level thinking or idea based discussions that demand greater communication with instructor and classmates. They assumed that "the advantage of distance education in delivering learning content in college level courses may not work as well for graduate level courses where more complex ideas are explored" (p. 43).

Discussion and Conclusion

Considering the positive trend for student achievement in distance education based on the recent reviews of literature, a higher achievement score in ODE settings was hypothesized in this study. However, the null hypothesis was not rejected: The result of this study indicated *no significant difference* in student achievement between ODE and F2FE (d = +0.023, k = 20). The result of non-significance was different when the studies were grouped according to Yes- or No-pre-test. In other words, the students in ODE significantly outperformed those in the F2FE settings in Yes-pre-test group of studies (d = +0.211, k = 9). In No-pre-test group of studies, the student achievement was not significantly different between ODE and F2FE (d = -0.106 k = 11).

The result is noteworthy in relation to the issue of methodological weakness in the area of distance education research. Methodological flaws of experimental studies in terms of the overall quality, e.g., controlling extraneous variables, random sampling, validity and reliability, have been identified by quite a few researchers (e.g., Bernard et al., 2004; Phipps & Merisotis, 1999). Furthermore, in order to evaluate student achievement in particular, the gap between pre-test and post-test scores should be measured and compared. If students at one course demonstrated higher prior knowledge in a pre-test and achieved higher marks in final test scores than those at its counterpart course, it cannot be the course effect. Thus, experimental studies should control the input variable (prior knowledge) to compare the output variable (final achievement). We assert, therefore, that the result from the Yes-pre-test studies is more reliable than that from the No-pre-test studies.

Assuming no difference in prior knowledge of students between ODE and F2FE sampled in No-pre-test studies, we may question how a pre-test could be more positively related to student achievement in ODE than in F2FE. There exists a possibility that a pre-test works as a moderator affecting teaching and learning processes of ODE settings. For example, a pre-test might provide information for the online instructors to understand students' academic ability to adjust the level of difficulty of the course content and offer appropriate feedback. Unlike F2FE settings where instructors gather information about student levels of understanding from facial expressions with direct communication, one of the biggest limitations of ODE was identified as a lack of information about students (Harasim, 2003). With regards to a self-learning perspective, a pre-test could help students recognize their own academic strengths and weaknesses. As Kozma (1994, p. 8) stated, "learning is an active, constructive, cognitive and social process by which the learner manages available cognitive, physical, and social resources to create new knowledge", the self-recognition might work more positively in ODE settings in terms of individual and participatory learning. The interpretations, however, are based on the assumption of equal variance of prior knowledge in no-pre-test studies. Whether a pre-test was a factor influencing more positively on student achievement in ODE or not cannot be asserted by this research. More experimental studies are called to prove the possibility that a pre-test may produce different outcomes in ODE and F2FE settings.

References

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

[1] Allen, M., Bourihis, J., Burrell, N., & Mabry, E. (2002). Comparing student satisfaction with distance education to traditional classrooms in higher education: A meta-analysis. The American Journal of Distance Education, 16(2), 83-97.

[2] Bates, A. W. T., & Poole, G. (2003).Effective teaching with technology in higher education Foundations for success. San Francisco: Jossey-Bass.

[3] Bates, A. W. T. (2000).Managing Technological Change: Strategies for College and University Leaders. San Francisco: Jossey-Bass.

[4] Bernard, R. M., Y. Lou, P. C. Abrami, L. Wozney, E. Borokhovski, P. Wallet, A. Wade, & M. Fiset. (2004). How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. Review of Educational Research 74(3): 379-439.

[5] Cavanaugh, C. S. (2001). The effectiveness of interactive distance education technologies in K-12 learning: A meta-analysis. International Journal of Educational Telecommunications 7(1): 73-88.

[6] Clark, R. E. (1983). Reconsidering research on learning from media. Review of Educational Research 53(4): 445-459.

[7] *Collins, M. (2000). Comparing web, correspondence and lecture versions of a second-year non-major biology course. British Journal of Educational Technology 31(1): 21-27.

[8] Cooper, H., and L. V. Hedges, eds. (1994).The handbook of research synthesis. New York: Russell Sage Foundation.

[9] Dutton, J., M. Dutton, & Perry, J. (2002). How do online students differ from lecture students. Journal of Asynchronous Learning Network 6(1). Available online at http://www.alnresearch.org/data_files /articles/full_text/6_1dutton.htm

[10] *Gainor, K. A., & Constantine, M. G. (2002). Multicultural group supervision: A comparison of in-person versus web-based formats. DE - age groups DE - counsellors - supervision of DE - student counselors so. ProfessionalSchool Counselling PD 6(2). Available online at http://www.findarticles.com /p/articles/mi_m0KOC/is_2_6/ai_96194760

[11] Glass, G. V. (1976). Primary, secondary and meta-analysis. Educational Researcher 5(10): 3-8.

[12] Harasim, L. (2003). The case for collaborative learning online. Educational Technology International 5(1): 77-100.

[13] Hedges, L. V., & Olkin,I.(1985).Statistical methods for meta-analysis. Orlando, FL: Academic Press.

[14] Hong, K. S. (2002). Relationships between students' and instructional variables with satisfaction and learning from a web-based course. Internet and Higher Education 5: 267-281.

[15] Jahng, N., & Krug, D. (2003). A review of issues affecting student satisfaction, learning, and dropout rates of students enrolled in online distance education. Paper presented at the Interdisciplinary Applications of Educational Technology Conference, November, Kyungbuk, Korea.

[16] *Johnson, M. (2002). Introductory biology "online": Assessing outcomes of two student populations. Journal of College Science Teaching 31(5): 312-317.

[17] *Johnson, S. D., Aragon, S. R., Shaik, N., & Palma-Rivas, N. (2000). Comparative analysis of learner satisfaction and learning outcomes in online and face-to-face learning environments. Journal of Interactive Learning Research 11(1): 29-49.

[18] Joy, E. H., & Garcia, F. E. (2000). Measuring learning effectiveness: A new look at no-significant-difference findings. Journal of Asynchronous Learning Network 4(1). Available online at http://www.aln.org/alnweb/journal/Vol4_issue1/joygarcia.htm

[19] *Kleinman, J., & Entin, E. B. (2002). Comparison of in-class and distance-learning students' performance and attitudes in an introductory computer science course. The Journal of Computing in Small Colleges, 17(6). Retrieved on February 11, 2004 from the Worldwide Web http://www.middlesex.mass.edu /carnegie/Carnegie_02/Attitude_Comparison.pdf

[20] Kozma, R. B. (1994). Will media influence learning? Reframing the debate. Educational Technology Research and Development 42(2): 7-19.

[21] *Leasure, A. R., Davis, L.,& Thievon, S. L. (2000). Comparison of student outcomes and preferences in a traditional vs. world wide web-based baccalaureate nursing research Course. Journal of Nursing Education 39(4): 149-154.

[22] Liao, Y. C. (1999). Effects of hypermedia on students' achievement: A meta-analysis. Journal of Educational Multimedia and Hypermedia 8(3): 255-277.

[23] Machtmes, K., & Asher, J. W. (2000). A meta-analysis of the effectiveness of telecourses in distance education. American Journal of Distance Education 14(1): 27-46.

[24] *Maki, W. S., & Maki, R. H.(2002). Multimedia comprehension ill predicts differential outcomes of web-based and lecture courses. Journal of Experimental Psychology: Applied 8(2): 85-98.

[25] Mayzer, R., & Dejong, C. (2003). Student satisfaction with distance education in a criminal justice graduate course. Journal of Criminal Justice Education 14(1): 37-52

[26] Murphy, T. H. (2000). An evaluation of a distance education course design for general soils. Journal of Agricultural Education 41(3): 103-113. Available online at http://pubs.aged.tamu.edu/jae/pdf/vol41 /41-03-103.pdf

[27] *Navarro, P., & Shoemaker, J. (1999). The power of cyber learning: An empirical test. Journal of Computing in Higher Education 11(1): 29-54.

[28] National Education Association (NEA).(2000). A survey of traditional and distance learning higher education members (Report). Washington DC. Available online at http://www2.nea.org/he/abouthe /images/dlstudy.pdf

[29] *Ocker, R. J., & Yaverbaum, G. J. (1999). Asynchronous computer-mediated communication versus face-to-face collaboration: Results on student learning, quality and satisfaction. Group Decision and Negotiation 8: 427-440.

[30] Olson, C. M., Rennie, D., Cook, D., Dickersin, K., Flanagin, A., Hogan, J. W., Zhu, Q., Reiling, J., & Pace, B. (2002). Publication Bias in Editorial Decision Making. The Journal of the American Medical Association (JAMA) 287(21): 2825-2828

[31] Olson, D. (2002). A comparison of online and lecture methods for delivering the CS 1 course. The Journal of Computing in Small Colleges 18(2): 57-63.

[32] Passerini, K., & Granger, M. J. (2000). A developmental model for distance learning using the Internet. Computers & Education 34(1): 1-15.

[33] Phipps, R., & Merisotis, J. (1999). The Institute for Higher Education Policy. What's the difference? A review of contemporary research on the effectiveness of distance learning in higher education. Washington, DC. Available online at http://www.ihep.com/Pubs/PDF/Difference.pdf

[34] Ramage, T. R. (2002). The "No significant difference" phenomenon: A literature review. e-Journal of Instructional Science and Technology 5(1). Available online at http://www.usq.edu.au/electpub/e-jist /docs/html2002/ramage.html

[35] Rovai, A. P. (2002). A preliminary look at the structural differences of higher education classroom communities in traditional and ALN courses. Journal of Asynchronous Learning Network, 6(1). Retrieved on February 7, 2004 from the World Wide Web http://www.alnresearch.org/data_files/articles/full_text /6_1rovai.htm

[36] Russell, T. L. (1999).The no significant difference phenomenon. NC: Office of Instructional Telecommunications, North Carolina State University.

[37] *Sankaran, S. R., Sankaran, D., & Bui, T. X. (2000). Effect of student attitude to course format on learning performance: An empirical study inn Web vs. Lecture instruction. Journal of Instructional Psychology 27(1): 66-73.

[38] *Schaik, P., Barker, P., & Beckstrand, S. (2003). A Comparison of On-Campus and Online Course Delivery Methods in Southern Nevada. Innovations in Education and Teaching International 40(1): 5-15.

[39] *Schoenfeld-Tacher, R., McConnell, S., & Graham, M. (2001). Do no harm-a comparison of the effects of on-line vs. traditional delivery media on a science course. Journal of Science Education and Technology 10(3): 257-265.

[40] *Schulman, A. H., & Sims, R. L. (1999). Learning in an online format versus an in-class format: An experimental study. T. H. E. Journal 26(11). Available online at http://www.thejournal.com/magazine /vault/A2090.cfm

[41] Shachar, M., & Neumann, Y. (2003). Differences between traditional and distance education academic performances: A meta-analytic approach. Available online at http://www.agecon.ksu.edu/accc/kcdc/PDF Files/differences.pdf

[42] *Smeaton, A. F., & Keogh, G. (1999). An analysis of the use of virtual delivery of undergraduate lectures. Computers & Education 32(1): 83-94.

[43] Spiceland, J. D., & Hawkins, C. P. (2002). The impact on learning of an asynchronous active learning course format. Journal of Asynchronous Learning Network 6(1). Available online at http://www.sloan-c.org/publications/jaln/v6n1/index.asp

[44] *Tesone, D. V., Alexakis, G., & Platt, A. (2003). Distance learning programs for non-traditional and traditional students in the business disciplines. Online Journal of Distance Learning Administration 6(4). Available online at http://www.westga.edu/~distance/ojdla/winter64/tesone64.html

[45] *Tucker, S. (2001). Distance education: Better, worse, or as good as traditional education? Online journal of distance learning administration 4(4). Available online at http://www.westga.edu/~distance /ojdla/winter44/tucker44.html

[46] Ungerleider, C., & Burns, T. (2003). A State of the Field Report to The Council of Ministers of Education, Canada and Industry Canada. A systematic review of the effectiveness and efficiency of Networked ICT in Education. Vancouver. Available online at http://www.cmec.ca/stats /SystematicReview2003.en.pdf

[47] *Wang, A. Y., &. Newlin, M. H. (2000). Characteristics of Students Who Enrol and Succeed in Psychology Web-based Classes. Journal of Educational Psychology 92(1): 137-143.

[48] *Waschull, S. B. (2001). The online delivery of psychology courses: Attrition, performance, and evaluation. Teaching of Psychology 28(2): 143-146.

[49] *Weems, G. H. (2002). Comparison of Beginning Algebra Taught Onsite versus Online. Journal of Developmental Education 26(1): 10-18.

[50] *Wegner, S. B., Holloway, K. C.,& Garton, E. M. (1999). The effects of internet-based instruction on student learning. Journal of Asynchronous Learning Network 3(2). Available online at http://www.sloan-c.org/publications/jaln/v3n2/pdf/v3n2_wegner.pdf

[51] Zhao, Y., Lei, J., Yan, B.,& Tan, S. (2004).What makes the difference? A practical analysis of research on the effectiveness of distance education. Available online at http://ott.educ.msu.edu/literature /report.pdf