EXAMINING THE EFFECTIVENESS OF DISTANCE EDUCATION: RESULTS FROM MULTI-LEVEL MODELLING

Tim Seifert, Bruce Sheppard, Ann Marie Vaughn, Memorial University, Canada

The meteoric rise of information technology as a means of providing educational opportunities has resulted in a transformation of educational institutions such that distance education (DE) has moved from the periphery to the mainstream (Merisotis & Phipps, 1999; Abrami et al., 2006). A review of the history of DE illustrates this transformation. Educational institutions first offered courses and programmes by correspondence as early as the 1800s, with the creation of the Chautauqua Correspondence College and by the 1950s 60 American universities had departments of correspondence courses (Moore, 2003). As the technology has changed new forms of DE evolved. While DE was text based in its early stages, radio and television created new opportunities for DE, which included the formation of public broadcasting (Moore, 2003). The intention of this "movement in the United States was to economize on teaching resources and subject matter expertise by distributing live lectures" (Bernard et al., 2004). This eventually gave way to tele-learning, and on-line learning which can include both the delivery of content and access to University materials. Concomitant to the formation of University departments responsible for the delivery of DE were the creation of Universities with entire programmes on-line (Moore, 2003), with a commitment from governments to the development and implementation of on-line learning (Council of Ministers of Education, Canada, 2001).

Although there is no question about its proliferation, how effective is DE? Several meta-analyses have been conducted with the intention of determining whether or not DE is as effective as classroom learning (CL). Beginning with the meta-analysis by Russell (1999), these meta-analyses have typically compared DE to CL on measures of achievement with the consistent finding that there were no overall differences in achievement, leading to the formation of the *no significant hypothesis* (NSD; Russell, 1999).

Following Russell (1999/2001), Bernard et al. (2004) found no difference in achievement outcomes between DE and CL (g=.02, k=318, N=54,775). Likewise, Ungerleider and Burns (2003) reported no differences between DE and CL on measures of achievement (g=0, k=12, N=1324), as did Cavanaugh, Gillan, Kromrey, Hess & Blomeyer (2004) in their meta-analysis of DE use in grades K-12 (g=-.028, k=14, N=7561). However, Cavanaugh (2001) reported a small effect (g=.15) favouring DE over CL in K-12 programmes.

Although there were no differences in achievement outcomes, these meta-analyses have yielded differences in attitude measures, albeit the effects were small. Bernard et al. (2004) reported a small difference (g=-.08, k=154, N=21,047) on attitude outcomes, with students in CL classes reporting slightly more positive attitude scores. Allen, Bourhis, Burrell & Mabry (2002) found that, on average, levels of students' satisfaction were higher in CL courses (d=.18, k=22, N=3866), but the size of the effect was somewhat small. However, that effect depended on the format of the course. When the DE course was correspondence-based, the size of the difference increased dramatically (r=.51, k=4, N=255). But when the DE course involved the use of video as the means of communication, the effect was small (d=.09, k=20, N=3483).

While the results of meta-analyses consistently support the *NSD* hypothesis, it is important to note that, in most cases, the effect sizes exhibited considerable heterogeneity. For example, effect sizes for the achievement outcomes in the study by Bernard et al. (2004) ranged from -1.25 to 1.25, and were roughly normally distributed. Likewise, Cavanaugh (2001), Ungerleider and Burns (2003) and Cavanaugh, Gillan Kromrey, Hess & Blomeyer (2004) reported heterogeneity of achievement effect sizes. Allen, Bourhis, Burrell & Mabry (2002) found the effect sizes in their meta-analysis of satisfaction were heterogeneous. Bernard et al. (2004) reported in their meta-analysis that the effect sizes for attitude outcomes were heterogeneous, ranging from -1.38 to 1.38 and approximating a normal distribution.

The fact that the effect sizes were heterogeneous and roughly normally distributed about zero calls into question the *NSD* hypothesis. These findings mean that the differences between DE and CL were greater in some studies than others, with students in DE performing better than those in CL approximately half of the time. The finding of *NSD* is not a consequence of performance in DE being comparable to that in CL, but rather students in DE outperforming those in CL in half of the studies and students in CL outperforming those in DE in the other half, resulting in a net gain of zero.

Our purpose in this study was to supplement the findings of the meta-analyses by continuing to explore the NSD hypothesis by examining the effectiveness of DE in a university setting. In doing so, we compared DE to CL in many courses, with many different instructors, and several years of implementation. In effect, our design mimicked eight meta-analyses of approximately 250 two-group, post-test only quasi-experiments. Given the diverse nature of instructors, courses, pedagogies, and students, a finding of no difference with minimal variation would be an important indication of the robustness of the NSD hypothesis.

Method

Data for this study came from student records provided by the registrar's office for the years 1999 to 2006. Variables included in the dataset were student identification numbers, courses taken, including those dropped, a distance education indicator for each course section, course instructor, and course grade. A number of restrictions were placed on the selection criteria for analyses: only undergraduate courses were examined, thereby by restricting the age range; a student could appear once per year (if they appeared more than once, a random selection for inclusion was made), and the minimum class size was set at 10. We selected only those courses that had been taught in both DE and CL formats in a single semester by the same instructor. The final result was a data set comprised of 39,689 course registrations, 61 different instructors teaching 47 different courses. The data were hierarchical in nature: students were nested within instructors. The sample sizes for instructors and students are presented in Table 1.

Table 1 Number of instructors and students in each year

	Analysis of average grade		Analysis of probability of not finishing the course	
Year	Number of instructors	Number of students	Number of instructors	Number of students
1999	29	4923	31	7704
2000	38	5585	39	7322
2001	45	6218	45	6851
2002	31	5504	33	6088
2003	33	5265	33	5658
2004	31	4983	31	5231
2005	31	3831	31	4055
2006	30	3380	30	3590

While many factors contribute to effectiveness and its definition, effectiveness was operationalized in two ways in this study: students' course grade was used as a measure of learning and finishing the course was used a proxy for satisfaction. A student was said to have finished the course if she received a grade greater than 20. Students were considered to have not finished the course if they dropped the course or received a grade less than 20. We rationalized using grades less than 20 as not finishing the course because such a grade indicates that a student receiving such a grade probably did not do the work, did not submit assignments or study for examinations.

A number of issues concerning validity can be raised when trying to compare distance education (DE) to on-campus (CL) formats. One obvious concern is whether or not a DE course is the same as its on-campus counterpart. Given the nature of DE, it seems reasonable to say that it is not the same course because of difference is interactions, opportunities for feedback, and access to resources, for example. While we acknowledge the differences, the question before us is not a necessarily a casual question to be answered in an experimental or quasi-experimental design. That is, any differences that might exist in students' grades that might exist between DE and CL classes my not necessarily be attributed to delivery format alone. For example, there

may be important differences in characteristics of students choosing DE rather than CL, such as age, gender, employment status, and reason for taking course, which may result in variations in students' motivation and performance. This leads to the conclusion that DE is a type of learning experience, and it is that learning experience that is being examined.

Results

Data for each year (1999-2006) were analyzed separately as a set of multi-level models using MPlus version 5.0 (Muthén & Muthén, 2006), and the results are presented in Table 2. Because the predictor variable in the model was a dummy coded variable (CL=0 and DE=1), the intercept value is the mean grade for CL, and the slope represents the difference in average grades between DE and CL formats. A positive slope would indicate that DE grades were higher than CL grades; a negative slope would mean that DE grades were lower than CL grades.

The average slope is an overall test of the *NSD* hypothesis, and there were no statistically detectable effects in seven of the eight years examined. In the year in which there was a statistically non-zero effect, the actual difference, although moderately small (ES=-.22), favoured CL courses. This finding is consistent with previous meta-analyses, and, on the face of it, seems to support the *NSD*. Although the overall effects showed no difference between DE and CL, the non-zero variances of the slopes indicate that there are considerable differences between instructors. That is, the average grade may not necessarily differ between DE and CL classes, but the results may depend on the instructor or course, and the values of the intraclass correlations indicate that differences instructors account for a significant portion of variance in grade.

A graphical representation of these differences for the year 2001, the year with the greatest number of instructors and students, may be found in Figure 1. The average grades in DE and CL courses were not different for some instructors. However, for some instructors, DE grades were higher than CL courses; for others DE grades were lower than CL courses.

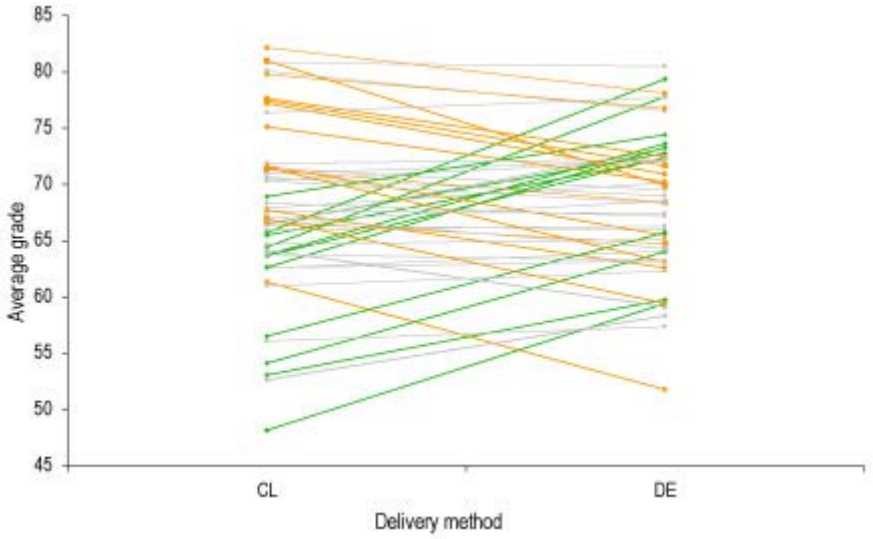

Figure 1 Average grade by delivery format and instructor: 2001

Table 2 Results from multilevel modelling of average grades in DE and CL classes

		Intercept (Average CL grade)		Slope (DE effect)	
Year	ρ	Mean	Variance	Mean	Variance
1999	.24	65.99*	35.87*	-.03	11.03*
2000	.39	65.90*	28.20*	.51	18.59*
2001	.34	66.66*	43.06*	1.63	21.98*
2002	.22	67.53*	28.80*	-.31	8.19*
2003	.37	68.88*	30.52*	.59	20.37*
2004	.49	69.07*	35.32*	-.85	23.62*
2005	.20	67.29*	24.45*	-1.78*	13.20
2006	.20	67.73*	38.00*	-1.04	9.72*

* denotes non-zero intercepts, slopes, and variances at p<.05. Student grades are level one units and instructors are level two units.

Similar to the modelling of course grade, differences in rates of not finishing a course were examined using a multilevel model with students' course status (finished or did not finish) and course type as level one variables nested within instructors (level two). The results, shown in Table 4, indicate some interesting points. The statistically detectable slope effects suggest that rates of not finishing a course are higher in DE than CL courses, and this appears to be fairly constant across years. It also appears that the rates of not finishing a course were fairly constant across courses, as suggested by variances in the slope that were not statistically different from zero.

Figure 2 is a graphical representation of the multilevel model results for the year 2001, the year with the largest number of instructors and students. As indicated in Table 4, the most slopes point to greater non-completion rate for students in DE than CL (a positive, non-zero average slope), and a large number of slopes appear to approximately parallel (little variation in slopes).

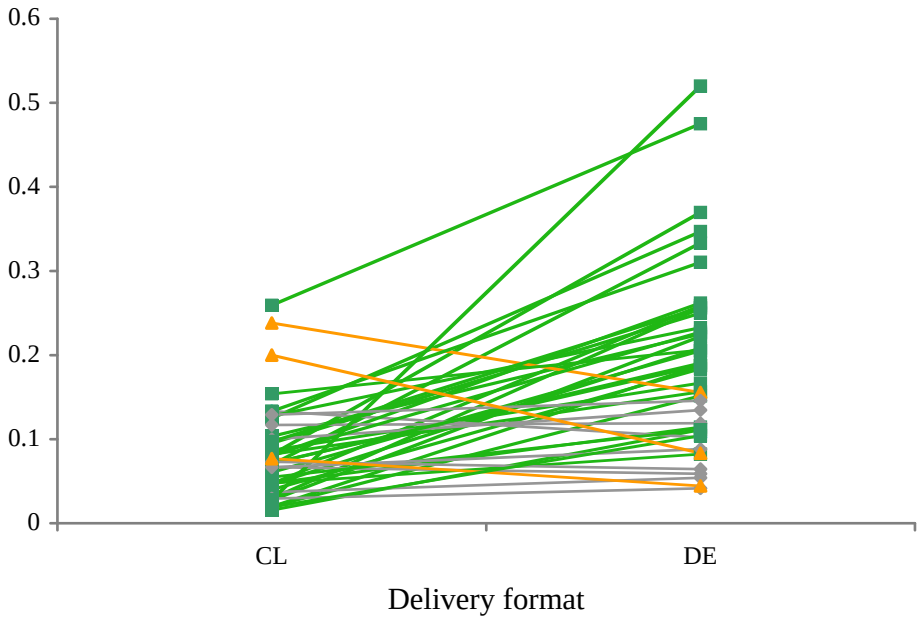

Figure 2 Probability of not finishing a course by delivery format and instructor: 2001

Table 3 Results from a multilevel model: Probability of not finishing the course by delivery format

	Intercept (CL)		Slope (DE effect)	
Year	Mean	Variance	Mean	Variance
1999	2.87*	.20*	1.06*	.01
2000	2.56*	.25*	1.01*	.03
2001	2.68*	.24*	1.04*	.17
2002	2.98*	.27*	1.27*	1.19
2003	2.97*	.25*	1.02*	.01
2004	3.28*	.22*	.92*	.02
2005	3.41*	.16*	1.18*	.03
2006	3.42*	.09	1.11*	.10

Note: The actual probability is given by the expression $P= \frac{1}{1+e^{(-a-b*x)}}$

Discussion

This study advances our understanding of the effectiveness of DE by scrutinizing the *NSD* hypothesis more closely. On average, there is no difference in student grades between CL and DE courses. However, the results clearly suggest that the *NSD* hypothesis cannot be taken at face value. There is substantial variation across instructors using DE and CL for course delivery and these findings are consistent with results from previous meta-analyses (Bernard et al., 2004; Cavanaugh, Gillan, Kromrey, Hess & Blomeyer, 2004; Ungerleider & Burns, 2003). While there is no difference on average, there exists considerable heterogeneity that needs to be investigated more thoroughly. What are the features of DE that can lead to effective learning? Unfortunately these data do not allow us to probe that question more deeply.

The analyses of rates of not finishing a course were more conclusive than those for grades. While the variances of slopes, representing variability in the DE effect, were not statistically different from zero, the average slopes showed a DE effect whereby students in DE exhibited higher rates of not finishing than CL students. Students in DE are much more likely than those in CL to not finish the course. Given that previous meta-analyses have suggested less student satisfaction in DE, and that students have freedom to drop and add courses, it may be the case that students' have a preference for CL. It may also be the case that the motivating factors of taking courses by DE such as complexity in time and place reduce the time available for students to be active participants in their learning and thus impact completion rates. Finally, student ability to learn independently impacts success in DE as does perhaps a perception that DE is not as demanding as CL. It may also be the case that the motivating factors of taking courses by DE such as complexity in time and place reduce the time available for students to be active participants in their learning and thus impact completion rates. Finally, students' abilities to learn independently impact success in distance learning, as might a perception that DE is not as demanding as classroom based learning.

Casual statements about the delivery formats in-and-of themselves are difficult to make because of threats to validity inherent in the study. Within any single comparison of DE and CL versions of a course there will be a number of factors that could account for any differences that might be observed (e.g., instructors, student characteristics, course requirements). While researchers may strive to achieve experimental control in their studies of DE and CL, it seems likely that such controls do not create circumstances mimicking those encountered in the field. Random selection or assignment is desirable, but is not the basis for course selection; course sections are taught by different instructors; content and pedagogy may differ between DE and CL courses. If so, the question is not about the delivery method per se, but rather, is the learning experience of students in DE comparable to that in CL. The answer from this study is that it can be, but isn't always. Subsequently, the question becomes when can DE work, and when does it not.

References

1. ABRAMI, P. C., BERNARD, R. M., WADE, A., SCHMID, R. F., BOROKHOVSKI, E.,TAMIM, R., et al. (2006). A review of e-learning in Canada: A rough sketch of the evidence, gaps and promising directions. Canadian Journal of Learning and Technology, 32, 3.
2. ALLEN, M., BOURHIS, 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, 83-97.
3. BERNARD, R., ABRAMI, P., LOU, Y., BOROKHOVSKI, E., WADE, A., WOZNEY, L., WALLET, P., FISET, M. & HUANG, B. (2004). How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. Review of Educational Research, 74, 379-439.
4. CAVANAUGH, C. (2001). The effectiveness of interactive distance education technologies in K-12 learning: A meta-analysis. International Journal of Educational Telecommunications, 7, 73-88.
5. CAVANAUGH, C., GILLAN, K., KROMREY, J., HESS, M. & BLOMEYER, R. (2004). *The effects of distance education on K-12 outcomes: A meta-analysis*. Naperville, IL: Learning Point Associates.
6. COUNCIL OF MINISTERS OF EDUCATION, CANADA (2001). *The e-learning evolution in colleges and universities*. Ottawa, ON: Government of Canada.
7. MERISOTIS, J. & PHIPPS, R. (1999). What's the difference? A review of contemporary research on the effectiveness of distance learning in higher education. Change, 31, 3.
8. MOORE, M. (2003). *From Chautauqua to the virtual university: A century of distance education in the United States*. Washington, DC: Office of Educational Research and Improvement.
9. MUTHÉN, L. & MUTHÉN, B. (1998-2006). *MPlus user's guide: Fourth edition*. Los Angeles, CA: Muthén & Muthén.
10. RUSSELL, T; L. (1999). *The no significant difference phenomenon*. Chapel Hill: Office of Instructional Telecommunications, University of North Carolina.
11. UNGERLEIDER, C. & BURNS, T. (2003). *A systematic review of the effectiveness and efficiency of networked ICT in education.* A report to the Council of Ministers of Education, Canada and Industry Canada.