Critical Issues in Dental Education

Estimating Transfer of Learning for Self-Instructional Packages Across Dental Schools

David W. Chambers, Ed.M., M.B.A., Ph.D.

Abstract: The most common topic of research in dental education is assessing the effectiveness of self-instructional units in various formats compared to lectures covering the same material. Generally, these studies are of high methodological quality and reveal mixed results or results slightly favoring self-instruction. All such studies, save one, have been conducted in the context of various single schools, thus confounding the effects of self-instructional format with factors particular to schools and their students. A reanalysis, using Cronbach’s generalizability analysis, was performed on a study in the literature that was conducted at six schools and measured student aptitude. The reanalysis found that the largest source of variance on immediate post-test quizzes for knowledge following a three-hour unit on disturbances in tooth development was the school at which the study was conducted (24 percent), followed by student aptitude measured by DAT score (20 percent). Difference in format among lecture, booklet, and audiotape presentations accounted for 5 percent of the variance. This reanalysis demonstrated that statistically significant results from rigorous experimental designs can overrepresent what is revealed by such research. The context-specificity of educational innovations may be underestimated because few studies are replicated across schools. Studies conducted as single schools, regardless of their methodological rigor, fail to address issues associated with potential transfer of findings to other schools.

Dr. Chambers is Professor and Associate Dean for Academic Affairs and Scholarship, Department for Academic Affairs, University of the Pacific Arthur A. Dugoni School of Dentistry. Direct correspondence to him at University of the Pacific Arthur A. Dugoni School of Dentistry, 2155 Webster Street, San Francisco, CA 94115; 415-929-6437 phone; 415-929-6654 fax; dchambers@pacific.edu.

Key words: self-instruction, generalizability, transfer, dentistry

Submitted for publication 7/28/05; accepted 11/2/05.

Surely, one of the intentions in publishing research in dental education literature is to encourage transfer of those innovations found to be effective to other schools. It is hoped, for example, that new educational methods that work in one school would be effective in others. The logic is patterned after experiments in the natural sciences or product testing. Some outcome associated with an intervention of interest is compared to a standard, and if found significantly different, the intervention is thought to be generally useful.

This article will raise two questions about this approach: 1) can we improve on the dichotomous (sig/NS) outcome of studies with some sort of quantitative measure of the impact of innovations? and 2) can the reports of our research contain more information about how representative findings in one school might be for dental schools generally? The change in perspective is one from “proving” something to “understanding” it.

Perhaps this analogy would help. It has been speculated in the news recently (hypothesized) that increases in gasoline prices are the result of gouging strategies by refiners (such as closing facilities for “repairs”), the accumulated results of overly restrictive environmental policies, speculation in commodities markets, a “risk premium” added because of geopolitical uncertainties, increased lavish consumption in the American lifestyle, and economic growth in Asia. It would be possible to assemble historical data to test each of these hypotheses separately, and it might be the case that any individually contributes a “statistically significant” amount to increased pump prices. But prudent policy would want to be informed based on the combined effects of all reasonable factors since some may be significant but trivial while others may kick in only when other circumstances prevail.

The most studied theme in the dental education literature involves self-instructional packages. Between 1970 and 2004, there have been fifty-seven articles published in the Journal of Dental Education investigating potential educational advantages of paper, video, slide-sound, computer-mediated, and
other self-contained units of instruction as a potential substitute for lecture or seminar presentations, in whole or as part of a course. Twenty-three of these studies are randomized trials, with pre- and post-tests.

Williams reviewed twenty-four dental education studies on self-instruction in various formats published between 1960 and 1980. These included six programmed texts, three teaching machines, seven slide-tape modules, two computer-assisted programs, three slide-guide packages, and two with mixed format. He concluded that these early self-instructional models claimed, in comparison to lectures on the same material, improved knowledge in about half the cases, less learning time in most cases, greater student satisfaction in all cases, and no superiority when comparing long-term retention or improved technical performance.

Dacanay and Cohen reported a meta-analysis eleven years later. This is one of the only meta-analyses to be published in the Journal of Dental Education. Thirty-four studies were found to exhibit sufficient mythological rigor. On post-tests, slightly fewer than half of the comparisons favored self-instruction, three favored the lecture format, and the measure of effect was .37 in the direction of self-instruction. The average reported study time for self-instruction was 77 percent of lecture time. (It is not customary to report attendance in lectures in this type of research, although some such correction appears appropriate.)

The most recent review of the impact of self-instruction in dental education is a 2003 article by Rosenberg et al. Only articles reporting studies of computer-assisted instruction format that used rigorous research designs were included in this review. Twelve studies were identified that met the established standards. In four of the twelve, the CAI format produced superior post-test scores; one was mixed; and one found superior scores for a seminar control group. Four of nine studies reported greater student satisfaction with the CAI format.

Although these studies support an expectation that self-instructional units might be effective replacements for parts of traditional instruction in general, there is no way to estimate from this body of research how likely the results are to transfer from one school to another or which school characteristics favor effective self-instruction. With a single exception, there has been no research in which the same self-instructional material was studied in more than one school. Our knowledge of the effectiveness of self-instruction in dental education is limited to format bound in unique contexts. We do not know how these results may travel.

In other studies, Kress et al. investigated the impact of Project ACORDE, a package of preclinical instructional materials in operative dentistry developed by a consortium of schools in the 1970s. The focus of their review was on adoption of the materials, not their effectiveness. The only research on self-instruction in dental education in which the same material was used at more than one dental school was reported as a series of four articles by Emling and Gellin.

This article presents a reanalysis of the data originally reported in the Journal of Dental Education in 1976. It demonstrates how a statistical procedure, generalizability analysis, can be used to understand research results at a deeper level and can provide estimates of how effects of self-instruction reported in one dental school might transfer to other schools.

Materials and Methods

The research reported here is a reanalysis of work conducted in the 1970s by Robert Emling and Milton Gellin. In the original study, the authors prepared a fifty-five-page booklet with behavioral objectives, glossary, information, pictures, radiographs, and practice cycles on “abnormal development of the dentition: developmental disturbances in the number of teeth.” The programmed text was converted to a three-hour lecture sequence and to an audiotape presentation, supported by reproduced visuals. Thus, the same material was developed in three formats. This pediatric dentistry unit was presented to students assigned at random to each of the instructional formats at six different dental schools. Faculty members were contacted at ten schools, and the six selected for participation included diversity in public or private support, size, and geography. In addition, the academic average for the DAT was available for each student participant.

A twenty-item multiple-choice test, with test-retest reliability of .71, was administered at the end of the lecture series or on request to students in the two independent study formats (the post-test). Approximately one month following the first test, the same examination was readministered unannounced as a measure of retention (the retention test). Although the data Emling and Gellin collected allowed
multiple characterization of student performance, the original analyses were all univariate ANOVAs and were spread across publications. There was an attempt to explore the interactions of school and format and of DAT aptitude and format, but this was done by considering each of the eighteen or nine combinations as classifications that were tested using a one-way ANOVA. This is not a recommended analysis approach, and the results could not be interpreted by Emling and Gellin. Thus there was no reporting of the relative effect of format, school, or student aptitude or the possible interaction of these factors.

Using the original dataset for reanalysis, the results of the original, univariate analyses were confirmed. In addition, a three-way ANOVA (school, format, aptitude) was performed for the post-test and retention test results. Finally, variance components were calculated for the test and retest data using the methods presented in Cronbach et al.’s The Dependability of Behavioral Measures.9

Results

The univariate analyses were in line with the general trend of self-instruction studies in dental education. For example, a one-way ANOVA on the post-test scores revealed a difference among formats that was significant at p<.05. The mean scores on the twenty-item post-test were 14.2 for the programmed text, 13.9 for the tape version of the text, and 13.3 for the lecture format of the same material. A separate one-way ANOVA was performed to test for differences across schools for post-test scores. A significance level of p<.001 was found, with average scores, combining across all three formats, ranging from 13.00 to 15.58. The retention test data were analyzed only for differences across formats. The one-way ANOVA failed to reach significance, with scores of 12.9 for the programmed text, 12.3 for the slide/tape version, and 12.2 for the lecture format. Although Emling and Gellin noted that student performance appeared to vary by format in different ways at some schools (an interaction effect) and that there were differences in DAT academic average across schools, these differences were not presented formally in their articles. The authors conclude that the results “demonstrate that the formats are, in the end, of similar effectiveness. Therefore, the choice of teaching format in a dental school can safely be based on considerations other than the amount of learning that will occur” (p. 76).5

The first set of reanalyses that were conducted involved a more comprehensive test of the significance of the three factors studied and of their interactions. The research design was a three-factor (school, format, aptitude), fully crossed design (all formats and aptitude levels at each school, all aptitude levels and school in each format, and all schools and formats in each aptitude level). The schools and aptitudes can be regarded as random, but the three formats constitute a fixed factor (specific formats are chosen, and the results will apply only to these formats).

A three-factor analysis of variance is appropriate for this design. These analyses are shown for the post-test (Table 1) and for the retention test (Table 2) and presented graphically in Figures 1 and 2.

Discussion

Generalizability Interpretation of Findings

It will be noted that these reanalyses confirm the analyses performed by Emling and Gellin: there are differences attributable to format in the post-test but not in the retention test, and there are differences across schools in the post-test situation. In addition, it emerges that student aptitude, as measured by the DAT academic average divided at quartile intervals, is also significantly associated with post-test and retention test scores. None of the two-way interactions are significant in this reanalysis.

The multifactorial analysis performed in the reanalysis offers a means for isolating variance components. This part of the procedure, known as generalizability analysis, begins by determining how much of the overall variation in results can be attributed to each measured source, including combinations of sources. These variance components have, among others, the following characteristics: 1) they are independent of sample size influence (the multifactorial ANOVA reported above is not free of these influences); 2) estimates of variance for main effects are corrected by removing interaction effects; and 3) the total variance can be expressed as percentages attributable to each source (totaling to 100 percent). Additionally, the estimated variance components are expressed in natural units, rather than “statistical”
units such as “sums of squares” or “mean squares.” In this case, the estimated variance components are points on the twenty-item quiz measuring mastery of material on abnormal development of the detention. In generalizability analysis, these elements are used to construct optimal measurement systems and to estimate error when measuring individuals. In this study, the estimated variance components will be used only to illustrate the relative magnitude of factors that affect post-test and retention test scores.

Estimated variance components and their percentage of the total variance in test scores are shown in the table below.

### Table 1. Initial post-test score analysis of variance and variance components on a unit in pediatric dentistry, classifying students by instructional format, school, and academic average on DAT

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Sig</th>
<th>Est Var</th>
<th>Var %</th>
</tr>
</thead>
<tbody>
<tr>
<td>School (S)</td>
<td>216.744</td>
<td>5</td>
<td>43.349</td>
<td>7.283</td>
<td>.001</td>
<td>2.996</td>
<td>24%</td>
</tr>
<tr>
<td>Format (F)</td>
<td>42.806</td>
<td>2</td>
<td>21.403</td>
<td>3.596</td>
<td>.03</td>
<td>.639</td>
<td>5%</td>
</tr>
<tr>
<td>Aptitude (A)</td>
<td>65.793</td>
<td>3</td>
<td>21.931</td>
<td>3.684</td>
<td>.01</td>
<td>2.467</td>
<td>20%</td>
</tr>
<tr>
<td>S x F</td>
<td>36.357</td>
<td>10</td>
<td>3.636</td>
<td>.611</td>
<td>.000</td>
<td>.000</td>
<td>0%</td>
</tr>
<tr>
<td>S x A</td>
<td>111.011</td>
<td>15</td>
<td>7.401</td>
<td>1.243</td>
<td>.446</td>
<td>.000</td>
<td>0%</td>
</tr>
<tr>
<td>F x A</td>
<td>10.900</td>
<td>6</td>
<td>1.817</td>
<td>.305</td>
<td>.000</td>
<td>.000</td>
<td>0%</td>
</tr>
<tr>
<td>Residual</td>
<td>2055.276</td>
<td>339</td>
<td>6.063</td>
<td>6.063</td>
<td>48%</td>
<td>6.063</td>
<td>5%</td>
</tr>
</tbody>
</table>

### Table 2. One-month retention test score analysis of variance and variance components on a unit in pediatric dentistry, classifying students by instructional format, school, and academic average on DAT

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>Sig</th>
<th>Est Var</th>
<th>Var %</th>
</tr>
</thead>
<tbody>
<tr>
<td>School (S)</td>
<td>485.250</td>
<td>5</td>
<td>97.050</td>
<td>13.243</td>
<td>.001</td>
<td>7.307</td>
<td>40%</td>
</tr>
<tr>
<td>Format (F)</td>
<td>22.148</td>
<td>2</td>
<td>11.074</td>
<td>1.511</td>
<td>.071</td>
<td>.071</td>
<td>0%</td>
</tr>
<tr>
<td>Aptitude (A)</td>
<td>79.926</td>
<td>3</td>
<td>26.642</td>
<td>3.635</td>
<td>.01</td>
<td>1.075</td>
<td>6%</td>
</tr>
<tr>
<td>S x F</td>
<td>93.692</td>
<td>10</td>
<td>9.369</td>
<td>1.278</td>
<td>.519</td>
<td>.519</td>
<td>3%</td>
</tr>
<tr>
<td>S x A</td>
<td>85.944</td>
<td>15</td>
<td>5.730</td>
<td>.782</td>
<td>.243</td>
<td>.243</td>
<td>13%</td>
</tr>
<tr>
<td>F x A</td>
<td>18.117</td>
<td>6</td>
<td>3.019</td>
<td>.412</td>
<td>.000</td>
<td>.000</td>
<td>0%</td>
</tr>
<tr>
<td>Residual</td>
<td>3292.105</td>
<td>339</td>
<td>7.293</td>
<td>7.293</td>
<td>39%</td>
<td>7.293</td>
<td>0%</td>
</tr>
</tbody>
</table>

It should also be noted that there is no presumption in the estimate of variance components regarding which sources should be regarded as effect and which are “error.” That is determined by the type of decision that users of the research wish to make. If this were a study of dental schools, the schools’ sources of variance would be a significant source of information, and format and aptitude would be “error.” If this were a study of student aptitude, school and format would become “error.” Only generalizability analysis permits estimates of sources of variation without predetermining what is to be labeled “error.”

Visual representations of the estimated components of variance are displayed in Figure 1. The overall area for each of the components of variance is approximately proportional to their contribution to overall variance in the scores on the initial and retention tests. In this graphic display, it is apparent...
Figure 1. Estimated proportion of variance in scores on post-instruction testing and retention testing in pediatric dentistry attributable to three instructional formats and four levels of student aptitude at six dental schools and to their interactions.

Figure 2. Retention test scores on a unit in pediatric dentistry for three instructional formats at six dental schools.
that the predominant source of variance is schools, including interaction effects involving schools. The variance contributed by presentation format is the smallest of the sources. Among the interactions, the one between school and aptitude for the retention test deserves comment. This interaction might be interpreted to mean that, among the schools in this study, there are differences in their views regarding whether students need to be told everything or whether they might be expected to learn it, or retain it, on their own—a teacher versus student orientation regarding who is responsible for learning.

There is also a very small school-by-format interaction that emerges in the generalizability analysis. This interaction was commented on at length in the Emling and Gellin articles. It was the case that the highest scores were achieved at the school where these two faculty members taught and that the highest set of scores was achieved by students at that school who studied the programmed text authored by one of the researchers. Undetected by the original authors, because they did not perform tests for interactions, is the trend for scores to be similar across formats at some schools and to be divergent at others. This effect is graphed in Figure 2. If the study had been conducted at schools B (the school of the original authors), E, or F, a one-way ANOVA on post-test scores would have supported the conclusion that the programmed text is a superior format; if the same study had been conducted at any of the other schools, the research would not have supported a difference in effectiveness.

Because the variance components generated in generalizability analysis are expressed in natural units (test questions in this case), it is possible to estimate confidence intervals for transfer to other schools from this type of research. This estimate is dependent on assumptions such as number of formats and students. A reasonable assumption would be that a single school uses a single format and has sixty students. The 95 percent confidence interval for the mean score on the post test would be 10.68 to 17.72 on the twenty-item test. Using the same assumptions, the 95 percent confidence interval on the mean score for the retention test would range from 9.34 to 16.46. The confidence intervals would be almost completely overlapping for the three instructional formats.

Conceptual Analysis of Findings

Reports of self-instructional formats that are effective at a single dental school may be misleading regarding how transferable these results would be to other schools. Statistical significance, even or perhaps especially in the case of rigorously designed studies, is not a sufficient justification for adopting research results from one context to another. This point has been made generally by Campbell and Stanley10 in their classic distinction between internal and external validity. High levels of methodological rigor in particular studies do not necessarily ensure high levels of transfer of these results to other contexts.

Research designs that test for the effect of a single causal factor can only answer a “yes or no” question: “Is there some context where it can be demonstrated that varying the factor in question makes a difference?” Regardless of the answer provided by the research, the questions remain: “Is the intended context for transfer sufficiently similar to the test context so that the results can be predicted?” and “Will the effect, even if positive, be noticeable when combined with the other sources of variance operating in the target context?” Generalizability analysis, especially the estimation of variance components, helps to address such questions.

It would be possible to read the Emling and Gellin study as supporting the use of programmed text as a more effective format for presenting a specific unit of instruction, at least on a test of immediate recall. This conclusion is supported by results at several schools and by combining results across the six schools studied generally. The difference between average scores using the programmed text format and the lecture method (14.2 minus 13.3) amounts to 4.5 percent of the possible points on the post-test. The difference between the two methods for retention testing was 2 percent of the total twenty points or 5 percent of the average on the initial test. By contrast, the difference between the best school (all formats combined) and the worst was 13 percent of the possible twenty points. Student aptitude, as measured by academic average on the DAT, also reveals differences of over 10 percent from the best-performing to the worst-performing groups.

It is possible to calculate estimates of variance components in the univariate case. It is not necessary to perform the full generalizability analysis; there are easy formulas for calculating omegasquared from z-values or eta-squared from R-values by removing the effects of sample size.11 Generally, such calculations reveal that variance components for measured effect can be very small, even for highly significant results.
Placing the impact of various sources of variance into relative context requires that they be measured and be reported in a fashion that makes them easy to grasp. The Emling and Gellin study is unusual in being conducted at multiple schools and further atypical in including at least one useful measure of a personal characteristic of learners (DAT score) that affects performance. The reanalysis reported here could not be conducted on the overwhelming majority of dental educational studies. The calculation of proportion of estimated variance attributable to each of the sources measured (the percentages appearing in the far right-hand columns of Tables 1 and 2 and shown graphically in Figures 1 and 2) can be readily understood. They lead to an intuitive grasp of how well a phenomenon is understood. They also point to those factors with the largest variance—those most in need of management through careful structure of circumstances or increasing numbers.

The analysis of interactions is possible only when multiple sources of variance are considered simultaneously and analyzed correctly. In the Emling and Gellin article, the authors were aware of the operation of interactions between school and format, although they could not quantify them adequately. Further work would need to be done before this kind of research would support reliable predictions of which schools would be good candidates for transfer. The authors did not report the names or otherwise characterize the successful and unsuccessful schools. The fact, demonstrated in the reanalysis, that the research would have been statistically significant had it been conducted at some of the schools but not significant at others should raise concerns about the great majority of the dental educational literature that reports results of studies conducted at single institutions. Because the authors did not use analytical tools that reveal interactions, they missed the larger and potentially interesting interaction between student aptitude and instructional format on retention of learned material.

**Conclusion**

Further research on the components and relative magnitude of factors affecting educational interventions across contexts appears promising. The following approaches appear to be warranted:

1. Where possible, more dental educational research should be conducted using a multifacto-