Title: From the Lab to the Classroom: Expanding and Scaling Up the Curriculum Domain
Abstract Body

Background/Context
Laboratory studies, including some small studies conducted in schools, often define a small, constrained (Paris, 2005) domain of content. For instance, in a small RCT of an online vocabulary instruction program, called the First 4000 Words, students studied 100 words, a small sample of the approximately 4000 words in the full curriculum (Fehr, Davison, Graves, Sales, Seipel, & Sekhran, in press). When the curriculum is so constrained, the outcome measure may contain a fairly large sample of the curriculum. In this laboratory study, the outcome measure contained 40 randomly sampled words from the 100 studied. Since the test contained one word for every 2.5 words in the curriculum, students would be expected to gain one point on the test for every 2.5 words learned. The content validity of the test was high in two senses: the test items were representative of the curriculum content and the test sampled a fairly large proportion (40%) of the words taught. Not surprisingly, Fehr et al. (in press (a)), found a large effect size roughly equal to one pre-test standard deviation.

Reading experts, however, argue that students must master somewhere in the neighborhood of 4000 Words in order to master the vocabulary represented in primary grade reading texts. Four thousand words is a far less constrained curriculum than the 100 words used in the laboratory study. Thus, vocabulary preparation for elementary reading means scaling up the curriculum in the laboratory setting to cover several thousand words, not just a small constrained sample of 100. In a second study lasting longer, Fehr, Davison, and Scholin (in press(b)) studied instruction of the full 4000 words. In this second study, the length of the instruction increased (from one to 13 weeks) but the material to be taught increased by a factor of 40, from 100 to 4000 words. The outcome test remained the same length, 40 words randomly sampled from the 4000 word curriculum. Since the test sampled 40 of 4000 words, it contained 1 word for every 100 on the test. As a result, students would be expected to gain 1 point on the test for every 100 words learned. In essence, from the first to the second study, the test became less sensitive to student learning. Because the scope of the curriculum changed in study 2, a student’s test would be expected to increase by one point for every 100 words learned rather than one point for every 2.5 words learned. The test remained content valid in the sense that it contained a representative sample of items in the learning domain, but it became a much smaller sample of the domain and hence less sensitive to student progress.

This paper considers the problem of designing a field study and the analysis of field study data given that the scope of the field study curriculum may be much broader than that in the laboratory research on which the larger field study is based. While the example is taken from vocabulary research designed to improve reading comprehension, similar challenges can arise in any subject area in which the curriculum is composed of discrete skills. Since vocabulary is a component of most academic instruction, the problems associated with teaching vocabulary apply broadly.

Purpose / Objective / Research Question / Focus of Study:
Moving from the laboratory to a more realistic classroom setting can mean a scaling up of the curriculum domain to the size expected to be covered in the classroom. As a result, an outcome measure of a given length may become less sensitive to student progress simply because the test contains a smaller sampling of the student outcomes.

An obvious solution would be to increase test length, but doing so is often not feasible within what we consider to be reasonable testing times. For instance, in the two studies cited above, the curriculum increased by a factor of 40. A corresponding increase in test length is not feasible.
Another obvious solution is to restrict ourselves to small, well defined curricular domains. However, the major outcomes defined by society tend to be broad and comprehensive; e.g. reading comprehension or mathematical achievement. This breadth is reflected in the NAEP assessments and the statewide tests employed in the NCLB school evaluations. If society is to become convinced that major achievement outcomes are improving or that achievement gaps are closing, the demonstrations seemingly must employ assessments covering broad outcomes.

The purpose of this presentation is to describe the problem of transitioning from laboratory results based on a constrained curriculum to the classroom in which there may a much broader list of desired outcomes. The presentation will discuss two issues, although the two issues do not constitute a comprehensive list.

The first issue involves estimating the length of intervention that would be needed to produce an effect of a given size. The amount of time needed would seem to depend on the rate of student learning and the percentage of items from the domain included in the outcome assessment. The implication is that as the domain size increases, assuming the outcome measure has the same number of items, the amount of time needed to produce an effect of a given size will increase. In short, the expected effect size for a treatment, and hence the power of the statistical test, will depend, in part, on an aspect of content validity, here called the sensitivity of the outcome measure and defined as the proportion of the content domain sampled by the assessment.

The second issue involves the data analysis. In the proposed analysis, the outcome variables are test items, not just total test scores, and the treatment effect is defined at the item level, rather than the total test level. This approach requires data on whether a particular item/objective in the outcome assessment had been covered in the instruction of a given student (i.e. 1 = taught, 0 = not taught) and whether the student had mastered the objective before the intervention began. Large domains can require some selectivity in the objectives covered. For instance, the teacher may individualize the instruction to focus on those items the student is least likely to know thereby omitting objectives likely to be known prior to instruction. Or the teacher may focus on just key objectives to make instruction manageable in the time available. Or the teacher may not reach some objectives in the time available.

Statistical, Measurement, or Econometric Model:

Content Validity and Effect Size. In large domains, student learning grows incrementally. For instance, if students in the treatment group are learning eight more words per week than students in the control group, then they will know 8 more words than the control group after week 1, 16 more after week 2, etc.

The observed effect, however, is only indirectly a function of the number of words learned. It is a direct function of the number of tested words learned. The expected number of tested words learned will depend on both the rate of learning and on the sensitivity of the test, here defined as the proportion of objectives in the domain contained on the test. If treatment students learn eight more words per week than control students, then one might expect treatment students to have learned 80 more words than control students at the end of 10 weeks. But if only 1% of the objectives are contained on the test, then after 10 weeks, one would expect less than one score point of difference (.01 * 80 = .8) on the test at the end of the 10 weeks.
The expected effect size then is a function of 3 factors: the rate of student learning (LR), the length of the intervention (T), and the sensitivity of the test (S). A simple estimate of the effect at the end of time T is: E(effect) = LR*T*S. Obviously, the expected effect size d can be obtained by dividing the expected effect by the population standard deviation or an estimate.

Usefulness / Applicability of Method:
Two problems arise in the application of this simple formula. First, one must estimate the student learning rate. This may be accomplished using data from a laboratory study or a small field study. For instance, in the Fehr et al. study in which each student participated for one week (5 classroom days), the difference between the treatment and control was approximately 8 words (7.8). With 1 word on the test for every 2.5 words in the curriculum, this would suggest an overall difference on all 100 words in the domain of about 20 words.

The second problem is to estimate the sensitivity of the test which depends on the length of the test and the size of the total domain. While in this particular example, the size of the domain is known, 4000 words, in most applications, the domain size is unclear. Methods of estimating it will be discussed. If students really can learn 20 more words per week and the test contains 1% of the items in the domain (40/4000 = .01), then the expected effect in test score units would be 3 score points at the end of 15 weeks. With a standard deviation of about 7 (6.7) on the pre-test in the laboratory study, this would translate into an effect of about .4 after fifteen weeks. By the end of a semester, one might expect an effect of about .4 assuming 15 weeks of effective instruction (even though a semester is 18 weeks).

Data Analysis:
In a large field study, two additional issues may arise. First, a given student may already know some of the content. One would not expect the intervention to improve scores on items whose answers were known in advance as evidenced by item performance on a pre-test. Second, some material may not be reached, either because the instruction was individualized or because some content could not be covered in the limited time available. Both of these conditions might suggest redesigning the analysis so that the main unit of analysis is the item, not just the total test. That is, these conditions might suggest a two-level analysis with items nested within persons. If \( \pi_{ij} \) is defined as the probability that person \( i \) answers item \( j \) correctly on the post-test, and the logit is defined as \( \eta_{ij} = \pi_{ij} / (1 - \pi_{ij}) \), then the level-1 model would be

\[
\eta_{ij} = \beta_{0i} + \beta_{1j} + \beta_{2j}y_{ij}
\]

and the level 2 model would be

\[
\beta_{0i} = w_0z_i + r_{0i}
\]

where \( \beta_{0i} \) is a random effect representing the achievement level of person \( i \), \( \beta_{1j} \) is a fixed effect representing the difficulty of item \( j \), \( y_{ij} \) is a predictor that indicates a treatment effect such that \( y_{ij} = 1 \) if item/objective \( j \) was taught to student \( i \) and unknown at pretest; \( y_{ij} = 0 \) either if the item/objective was not taught or the item/objective was known before the instruction as indicated by a pretest. The key null hypothesis in the level-1 model is \( \beta_{2j} = 0 \) indicating no effect of teaching an item that was previously unknown to the student. In the level-2 model, \( z_i = 1 \) if person \( i \) was randomly assigned to the treatment and \( z_i = 0 \) if person \( i \) is assigned to the control group. The simple treatment effect is tested with the null hypothesis \( w_0 = 0 \).
The analysis of the data that will be described will involve fitting several models that vary in whether they include the terms involving the level-1 treatment variable $y_{ij}$ and the level-2 treatment variable $z_i$. The interpretation of several possible outcomes will be discussed. For one of the interesting outcomes, is the level-2 treatment effect is not significant (with and without the level-1 treatment included) but the level-1 predictor is significant. In this case, the significant finding at level-1 suggests the treatment does improve outcomes of taught objectives not previously known by the student. Coupled with the nonsignificant finding for the level-2 effect, it suggests that the overall treatment effect may be weak either because many of the objectives were known prior to the intervention or because some objectives were not covered. In other words by fitting various models, one may be able to identify reasons why the overall treatment effect was not significant, some of which may be associated with scaling up the size of the curriculum (i.e. some material not taught).

The presentation will include results from the online vocabulary intervention in which the computer administration allowed monitoring of which specific objectives on the post-test were taught to a particular student and which were “skipped” due to the individualization of the instruction. Also, because of pre-testing, it was known which tested objectives were and were not known prior to instruction.

**Conclusions:**
For those moving from the lab to the classroom in which a much larger domain of material is covered, it will be recommended that researchers take into account the estimated rate of student learning and the sensitivity of the outcome measure to estimate the expected effect at various time points in the intervention. Outcome measures should be delayed until the expected effect has reached a detectable size. The major challenges with this approach involve estimating the rate of learning from pilot data and the sensitivity of the outcome measure, particularly the domain size on which the sensitivity index depends.

It will also be suggested that, where data are available on the instruction of particular outcomes and or the student’s prior knowledge of particular outcomes, researchers may want to adopt an evaluation of treatment effects employing the item as the basic unit of analysis. In effect, the item level model includes a variable which can be useful in estimating the effect of the treatment on outcomes given that the outcome is taught and not known prior to the intervention. The challenges include the need to have data on student prior knowledge at the item level and data on delivery of instruction at the outcome level.
Appendix A. Reference

Fehr, C., Davison, M. L., & Scholin, S. (in press(b)). The Effect of Online, Large-Scale Vocabulary Instruction on Listening and Reading Vocabulary. *Information Technology, Education, and Society*.
