Assessment Practices in Curriculum Materials Research: Do Students Learn from the Pretest?

Robert Ochsendorf
Curtis Pyke
The George Washington University

Paper presented at the Annual Meeting of the National Association for Research in Science Teaching
New Orleans, LA
April 15-18, 2007

This work was conducted by SCALE-uP: A collaboration between George Washington University and Montgomery County Public Schools (MD): Sharon Lynch, Joel Kuipers, Curtis Pyke, and Bonnie Hansen-Grafton, principal investigators. Funding for SCALE-uP is provided by the National Science Foundation, the U.S. Department of Education, and the National Institute of Health (REC-0228447). Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the position, policy, or endorsement of the funding agencies.
Abstract

This paper reports on a study of the effect of pretesting on middle-school students’ learning in different motion and forces curriculum conditions. Since current policy in science education calls for educators and students to view assessment situations as learning opportunities, pretests might be implicated as a confounding cause of learning in curriculum evaluation. Therefore, this research employs a four-group design to test the pretest effect hypothesis. Pretest classrooms were randomly selected from five matched pairs of schools (each pair implementing one of two curriculum conditions) (n=295 students) while students in the remaining classrooms did not receive pretests (n=1725 students). Findings reveal a correlation between pretest scores and posttest scores as expected, but no differential pretest effects on posttest performance across curriculum conditions. Overall, whether or not students took the pretest did not influence posttest scores. Implications for contemporary curriculum evaluation research are discussed.
Introduction

This paper addresses the effect of pretesting in contemporary science curriculum evaluation. Specifically, does administering a pretest influence future learning during a curriculum experiment and if so, do different curriculum conditions produce different pretest effects? Understanding pretest effects in curriculum experiments continues to be of interest because assessment’s role in curriculum and instruction is changing. In science education and the field of curriculum and instruction more broadly, embedded and formative assessments are being promoted as tools to enhance learning (Black & Wiliam, 1998; NRC, 1996). In fact, Black and Wiliam’s research elevated the status of formative assessment and redefined it as all those instructional activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged. If the collection of work on assessment has changed the culture and use of assessment in curriculum and instruction then it is reasonable to expect that the combination of a pre-test and a high quality curriculum might combine to artificially enhance the observed effects of a new curriculum innovation. Therefore, by explicitly studying the effect of pretests we hope to better understand the threats to valid research conclusion on one hand and the possibility in improving learning on the other.

To elaborate further on this research vs. practice dilemma, we begin with an observation that educators are clearly encouraged to incorporate assessment experiences into the learning process in a meaningful way, but note that curriculum evaluation researchers require objective and independent pretests to equate groups and isolate the effect of a curriculum intervention in their research. In terms of the latter, that pretesting becomes part of the experience of the research subjects producing unintended effects on
posttest scores and confounding attributions of improvement to the intervention is not disputed (Kavlem, Sundet, Rivo, Eilertsen, & Bakketeig, 1996; Sprangers & Hoogstraten, 1989). However, the significance of findings varies with prior investigations of pretest effects concluding that pretests effects are more prevalent in studies with affective-type outcomes (e.g. attitudes, beliefs, values). Relevant to the current work, no consistent pretest effects have been found when researchers use cognitive measures with a longer rather than shorter pretest-posttest window (e.g., months), and pretest studies to date have often used generalized achievement measures or IQ tests as the assessment (see for example, Welch & Walberg, 1970) than measures of the kind of complex concepts and big ideas targeted by curriculum units.

This current study was motivated by a real interest in how middle school students in the quasi-experiments of the Scaling Up Curriculum for Achievement Learning and Equity Project (SCALE-uP)(REF or web site) made use of their pretest experience and if there is any empirical evidence to be concerned that pretesting confounds inferences. However, the work of this paper became more than a report on a study of internal validity as we noticed that pretests and formative assessments frequently refer to similar activities occurring in classrooms and that it might not be enough simple to view learning from pretests as an unintended side effect, but an intended practice. As described in more detail below, the intervention of interest is a reform-based middle school science curriculum unit and the assessment of interest targets the big ideas of motion and forces.

Background

Scaling Up Curriculum for Achievement Learning and Equity Project

This research study was undertaken as part of a larger, multi-year quasi-experimental curriculum studies project (SCALE-uP) interested in the effectiveness and
scaling-up of middle school science curriculum materials. The curriculum materials were rated using the Project 2061 Curriculum Analysis (see Kesidou & Roseman, 2002 for more details) and the research has involved the implementation of curriculum materials in 38 middle schools in a large, diverse school district over the past 5 five years (2002-2007). The materials cover astronomy, physical science, and chemical science topics in grades 6 through 8. Previous results show some of the curriculum materials are more effective than standard materials being used in the schools (Lynch, Kuipers, Pyke, & Szesze, 2005) while other materials have proven less effective than standard materials (see Pyke, Lynch, Kuipers, Szesze, & Watson, 2005). Data from grade 6 students using the physical science unit *Exploring Motion and Forces* (Harvard-Smithsonian Center for Astrophysics, 2001) are analyzed here.

A non-experimental control group pre-post quasi-experimental methodology was used in all the SCALE-uP curriculum studies (Cook & Campbell, 1979). The design employed random selection of treatment and comparison schools from demographically similar group of schools. SCALE-uP divided the 38 middle schools in the large school district into five School Profile Categories (SPCs) primarily indexed by socioeconomic status (SES). A pair of middle schools was randomly selected from each SPC. One school from each pair of schools was randomly assigned to implement the treatment curriculum unit for two consecutive years. Taking steps to minimize the impact of the nesting of students in schools student level ANCOVA methods (using pretests scores as a control variable) were the primary statistical analyses conducted (see Pyke, Lynch, Kuipers, Szesze, & Watson, 2006). The use of Hierarchical Linear Modeling in some exploratory analyses has confirmed the robustness of the ANCOVA findings in terms of significance tests, but has revealed significant classroom level effects and increase in effect size.
estimates for the curriculum effect (Rethinam, Pyke, & Lynch, in press; Rethinam, Pyke, & Lynch, 2007). In addition to improving statistical analyses of the SCALE-uP data, the research plans for 2005-2006 SCALE-uP included tests to study potential threats to internal validity.

**Pretesting as a Threat to Internal Validity in Quasi-Experiments**

Pretests are used in educational research for a number of reasons. For example, pretest scores are used in assigning individuals to treatment and comparisons groups to establish initial group equivalence on a criterion of interest. Pretest scores are also used in calculating gain scores (posttest – pretest) as a way of studying value added. More analytically, when random assignment is not possible and there is no reason to expect initial group differences, pretest scores are used in data analyses to account for student differences that might arise due to sampling (e.g., ANCOVA). Clearly there are good reasons for researchers to consider pretesting in curriculum evaluation. However, in the research methodology literature pretests have been theorized to influence outcomes in a few major ways. First, students might learn directly from the pretest and this pre-test experience would lead to higher scores on the posttest. Second, during the pretest experience, sensitization might occur as students become “aware” of and scaffold the ideas to be learned. Third, Cook and Campbell (1979) point out that familiarity with a test can sometimes enhance performance because items and error responses are more likely to be remembered at later testing sessions, a particular concern when the same test is used at pre and post testing.

Learning from pretests, in particular differential learning across groups, is a realistic concern considering the design of the SCALE-uP studies. This concern falls under the scope of *internal validity*, or the trustworthiness of statements made about
whether the M&F curriculum causes differences in learning when compared to learning with standard curriculum. In general terms, threats to internal validity for the no-treatment pretest-posttest quasi-experimental design include: selection-maturation, instrumentation, regression to the mean, and local history (Cook & Campbell, 1979). Differential effects of pre-tests across condition might occur due to selection-maturation if one group is more ready than another to benefit from a scaffolding effect due to selection flaws, or can be considered under the category of local history. For example, instructional initiatives promoting use of formative assessment practices in select schools might result in differential attempts to explicitly learn from pretests. If selection-maturation and/or local history effects are present in only treatment or comparison schools, then inferences about curriculum effects are compromised. Research on the assessment of prior knowledge and its role in learning is relevant to understanding the potential of pretesting to threaten internal validity.

**Prior Knowledge and Learning**

Considerable research in the learning sciences has suggested that it is important to account for students’ prior knowledge for the purposes of instruction (National Research Council, 2000; 2005). The role of prior knowledge in the learning process has been well researched and the results of many studies are generally in agreement. A review of research studies on prior knowledge conducted by Dochy, Segers, and Buehl (1999), concluded that prior knowledge has a strong positive relationship with improved learning outcomes. That is, when students come to learning situations with more prior knowledge about a topic, they are likely to learn more than other students who possess less prior knowledge. When it comes to curriculum research it is important to consider and recognize the role that prior knowledge (often measured by pretests) plays in influencing
subsequent learning. With that in mind, a useful pretest should effectively measure prior knowledge, predict future learning, but should not interact with research interventions to influence outcomes.

Assessment and Learning

As noted in the introduction, prior research on pretest effects does not seem to directly address the concerns about threats to internal validity that would result from learning from the pretest assessment: Especially for this sample, context and type of assessment. Even though education and psychology researchers of the 1970’s and 1980’s seemingly converged to resolve the issue of pretest effects on learning outcomes (Welch & Walberg, 1970; Wilson & Putnam, 1982), we believe there is considerable reason to revisit this question today.

Related to increased awareness in the role of prior knowledge in learning has come a call for a change in assessment to better identify and use prior knowledge in instruction. In a report on the current role of assessment in education, the National Research Council (NRC) suggests that students need to be more engaged in monitoring their own learning and to be more self-directed (2001a) when it comes to assessments. They suggest that teachers ought to help students learn more from assessment situations. More specifically, the NRC also recommends that “science teachers must use assessment information to adapt their teaching to the needs of students and that students must be more involved in the assessment process” (2001b). The National Science Education Standards calls for more emphasis on students being engaged in ongoing assessment of their work and that of others (1996). And, the AAAS, Project 2061 placed emphasis on embedded assessment in putting forth its curriculum analysis procedures (Stern & Ahlgren, 2002)
The movement toward learning from assessment was further bolstered by empirical evidence from, Black and Wiliam (1998) in a review documenting the positive effects of formative assessment in learning. These researchers have called for students and educators to pay greater attention to assessment experiences prior to and during instruction. We believe this movement calling for increased attention to assessment situations increases the possibility that pretest experiences in SCALE-uP studies has acted as a learning opportunity in some classrooms. However, empirical research does show that pretest effects are more likely to be present when the dependent variable is a self-report measure (i.e., personality, attitude, opinion) and when the testing window is relatively short (hours, days or perhaps weeks). Pretest effects on outcomes of student learning have been markedly less prevalent when using achievement measures and longer testing windows. (Bracht & Glass, 1968; Welch & Walberg, 1970). In their meta-analysis of pretest effects in experimental designs, Wilson & Putnam (1982) recommended that researchers continue to use pretests but attempt to estimate their effects.

Despite these claims, we believe that recent trends in the educational climate in which assessments are operating provide reason to explore the impact of pretests on subsequent student learning. For this reason, we have designed this pretest effects study to determine if students’ experience in taking a pretest enhances their performance on the posttest and if there are differential effects across curriculum conditions.

Research Questions

Thus, this paper addresses the following research questions:

1) Is there a relationship between performance on the pretest and performance on the posttest?
2) Do students that take the pretest learn more than students that take the post-test only?
3) Does the students’ experience with the pretest interact with the experimental condition to influence student outcomes?

4) Does demographic subgroup (gender, SES, or ethnicity) determine the effect of a pretest?

5) Does the combination of demographic subgroup and pretest condition interact with curriculum condition?

Methods

This current study builds on the SCALE-uP research methodology by employing a Solomon four-group design (Campbell & Stanley, 1963) to isolate pretest effects. Pretests were administered to randomly selected intact classrooms of students (n=295 students) from five schools implementing *Motion and Forces* and from five schools implementing an alternative curriculum. The remaining students in the schools did not receive pretests (n=1725).

The treatment group of schools implemented *Exploring Motion and Forces* (Harvard-Smithsonian Center for Astrophysics, 2001), a sixth grade physical science curriculum unit. The comparison group was asked to implement traditional motion and forces materials that had been in use in the school district prior to the study. Pretests were administered by personnel from the district’s accountability office so that teachers would not be aware of the contents of the assessment and also so pretest effects could be isolated to a student variable. Pretesting was completed for all students in the same two-day window approximately two weeks prior to the beginning of the unit. The posttest was administered following instruction on motion and forces by the classroom teacher in all treatment and comparison classrooms. The pretest-posttest window ranged from four months to six months in both treatment and comparison conditions. The Motion and Forces Assessment (MFA) is a curriculum independent assessment composed of ten selected and constructed response items across four everyday contexts involving objects
in motion and the forces acting on them. The MFA, developed by SCALE-uP, was administered both before and after instruction to provide pre and post measures of student understanding. Cronbach’s alpha for the ten items of the MFA used in calculating the weighted score indicated the internal consistency of the assessment ($\alpha = 0.52$). The weighted score range is 0 to 100. A standard setting process was undertaken to determine cut scores that distinguish among four levels of understanding: flexible understanding (scores ranging from 70 to 100), some fluency with ideas (50-70), context-limited understanding (20-50), and no understanding (0-20). (For more information on the process of MFA development, see Pyke & Ochsendorf, 2007.)

All analyses were conducted using SPSS for Windows (Version 12.0) controlling for the unequal cell sizes.

**Results**

Consistent with the relationship of prior knowledge and outcomes in the literature, pretest scores (as indicators of prior knowledge) were moderately related to posttest performance, $r(293) = .38$, $p < .01$. However, while pretest scores were related to outcomes there was no statistical evidence of students learning from the pretest as students who took the pretest did not perform statistically different from students who did, $F(1, 2018) = 2.16$, $p > .05$. In addition, when assessing the interaction between pretest and experimental condition, the results of a two-way ANOVA suggest that experiencing the pretest does not interact with experimental condition to influence performance on the posttest with $F(1, 2016) = 0.314$, $p > .05$. Group means are presented in Tables 1 and 2.

In determining if pretest effects exist in different ways for different subgroups of students classified by SES, eligibility for Free and Reduced Meals Status (FARMS)
(Never Eligible and Ever Eligible) was used as a proxy for SES. In addition, tests for interaction effects were conducted to determine if gender or ethnicity had a differential impact on the pretest-posttest performance (Research Question 5). The test of the 3-way interaction revealed no significant interaction effect for FARMS, $F(1, 1980) = 2.90, p > .05$, gender, $F(1, 1980) = .56, p > .05$, or ethnicity, $F(3, 1972) = .95, p > .05$. Also, the test of the two-way interaction (Research Question 4) revealed no significant interaction effect for FARMS and curriculum condition, $F(1, 1982) = 2.95, p > .05$, gender and curriculum condition, $F(1, 1984) = 3.91, p > .05$, or ethnicity and curriculum condition, $F(3, 1972) = .03, p > .05$. Table 3 reports the subgroup means.

Conclusions

Results suggest that students’ learning outcomes are related to pretest scores indicating prior knowledge about force and motion ideas is related to subsequent learning. While the literature predicts this relationship between prior knowledge and outcomes, we were interested to know if pretesting caused learning and if there was an interaction of the pretest experience with the curriculum conditions. Finding no interaction and no main effect for pretesting is encouraging from a curriculum evaluation perspective; knowing that the pretest experience did not cause learning and/or advantage one group over the other reduces concern over a pretest threat to internal validity. The results are also consistent with prior studies suggesting that when using tests for non-affective variables with a longer pretest-posttest window there is not likely to be a pretest effect on outcomes.

While we are unaware of prior research that investigates how particular subgroups might be impacted by pretest experiences, this research suggests that the lack of a pretest effect is consistent for all sixth grade science students across levels of ethnicity, SES, and
gender. These findings occur with a testing window of a few months and with an assessment that is targeting specific science content standards. It is interesting to note that no specific subgroup of students seemed to leverage the pretest experience as a learning opportunity that was detectable by the MFA posttest. However, these findings are somewhat consistent with previous research indicating that longer testing windows are not likely to result in the presence of a pretest effect.

In the context of the formative assessment literature, this study indicates that sixth grade science students may not, on their own, make the most of assessments as learning opportunities. Perhaps, if provided more explicit opportunities to engage in assessments as learning opportunities and to receive feedback regarding those opportunities, these students could have demonstrated from such as assessment. Though, given the current emphasis on learning from assessment, one has to wonder why MFA pretests did not provide more of a learning opportunity for at least some of the students.

Suggestions for Future Research

We echo Wilson & Putnam (1982) in advising that when employing pre-post research designs in any content area and age level, pretest effects should be estimated to assess threats to internal validity. However, the field would be well served by an explicit research effort to determine under which conditions students do learn from pre-testing situations in ways that influence a given study’s research findings.

In addition, prior implementation studies by SCALE-uP have utilized classroom teachers in both the treatment and comparison groups to act as pretest administrators. As a way to eliminate the possibility of a teacher effect (i.e., teachers becoming sensitized to the content of the MFA and altering their classroom practice) in this pretest study, SCALE-uP arranged for other district employees to act as pretest administrators, thus
eliminating the possibility of a teacher pretest effect. Additional studies seeking to isolate a possible teacher effect relative to pretests would be encouraged to complement the results presented here. Such findings could have important implications for the methods by which researchers administer assessments to students. We would at this time have to recommend our procedures for reducing the teacher threat.

Finally, considering the increasing push for use of pretests and embedded assessment as an instructional strategy in curriculum materials, more research needs to be conducted to determine if, when, and what kind of pretest and curriculum conditions can and do influence learning. Our study clearly shows that just giving a pretest does not in and of itself have any instructional value added.
References


Table 1. Posttest scores of students who did and did not experience the pretest.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Posttest Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-1,C-1</td>
<td>295</td>
<td>Yes</td>
<td>Yes</td>
<td>51.81</td>
</tr>
<tr>
<td>T-2,C-2</td>
<td>1725</td>
<td>No</td>
<td>Yes</td>
<td>53.89</td>
</tr>
</tbody>
</table>

Table 2. Solomon four-group design.

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Pretest</th>
<th>Pretest Mean</th>
<th>Treatment</th>
<th>Posttest</th>
<th>Posttest Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-1</td>
<td>164</td>
<td>Yes</td>
<td>32.70</td>
<td>Experimental</td>
<td>Yes</td>
<td>53.87</td>
</tr>
<tr>
<td>T-2</td>
<td>790</td>
<td>No</td>
<td>-</td>
<td>Experimental</td>
<td>Yes</td>
<td>55.29</td>
</tr>
<tr>
<td>C-1</td>
<td>131</td>
<td>Yes</td>
<td>34.26</td>
<td>Comparison</td>
<td>Yes</td>
<td>49.22</td>
</tr>
<tr>
<td>C-2</td>
<td>935</td>
<td>No</td>
<td>-</td>
<td>Comparison</td>
<td>Yes</td>
<td>52.23</td>
</tr>
</tbody>
</table>
Table 3. Mean MFA posttest scores for subgroups of students that did and did not experience the MFA pretest.

<table>
<thead>
<tr>
<th>Student Subgroup</th>
<th>Not Pretested</th>
<th>Pretested</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Treatment</td>
<td>Comparison</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>58.47</td>
<td>56.51</td>
</tr>
<tr>
<td>Female</td>
<td>52.14</td>
<td>48.63</td>
</tr>
<tr>
<td>FARMS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>61.76</td>
<td>59.05</td>
</tr>
<tr>
<td>Ever</td>
<td>47.05</td>
<td>39.60</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>49.64</td>
<td>42.40</td>
</tr>
<tr>
<td>Asian American</td>
<td>55.35</td>
<td>60.12</td>
</tr>
<tr>
<td>Hispanic</td>
<td>47.26</td>
<td>37.90</td>
</tr>
<tr>
<td>White</td>
<td>62.53</td>
<td>58.01</td>
</tr>
</tbody>
</table>