Developing a conceptual assessment for a modular curriculum

Paula V. Engelhardt,[1] Steve Robinson,[1] Edward P. Price,[2] P. Sean Smith,[3] and Fred Goldberg[4]
[1]*Department of Physics, Tennessee Tech University, 110 University Drive, Cookeville, TN 38505*
[2]*Department of Physics, California State University San Marcos, 333 S. Twin Oaks Valley Rd., San Marcos, CA 92096*
[3]*Horizon Research, Inc., 326 Cloister Court, Chapel Hill, NC 27514*
[4]*Department of Physics, San Diego State University, 5500 Campanile Drive, San Diego, CA 92182*

This paper presents the development and field-testing of a multiple-choice content assessment designed to measure student learning gains of prospective elementary teachers enrolled in a course using the *Next Generation Physical Science and Everyday Thinking* curriculum over the course of one term. Preliminary results of the initial pilot in spring 2017 with a small group of experienced instructors and a larger second administration by members of the *Next Gen PET-Faculty Online Learning Community* in fall 2017 are presented. Practical aspects of developing and evaluating the effectiveness of the assessment instrument for a modular curriculum and administration by a large collaboration are discussed.

I. INTRODUCTION

The National Research Council's Framework for K-12 Science Education and Next Generation Science Standards (NGSS) [1] are the basis for K-12 science standards in many states. According to NSTA, "nearly two-thirds of U.S. students live in states that have education standards influenced by the Framework for K-12 Science Education and/or the Next Generation Science Standards" [2]. While the Framework and NGSS are intended for K-12 science education, university science courses that prepare future teachers should be consistent with the Framework and standards, including the integration of content, practices, and crosscutting concepts [3].

In response, the *Next Generation Physical Science and Everyday Thinking* (*Next Gen PET or NGP*) materials were developed based on *Physics and Everyday Thinking (PET)* and related curricula [4-6]. *Next Gen PET* is a flexible set of materials for university science courses for prospective elementary teachers that can support instructors and students in small or large enrollments and with physics or physical science content. It consists of nine stand-alone units: Magnetism, Static Electricity, Energy and Motion, Potential Energy and Fields, Interactions and Forces, Waves and Sound, Light, Physical Changes, and Chemical Reactions. Here we report on all but the Physical Changes and Chemical Reactions units, which will be presented elsewhere. The Interactions and Forces unit is about twice the length of the other units. Each unit has two versions, one for small, studio-style classes that are able to accommodate extensive small-group laboratory work and discussion, and one for large, lecture-style classes. Both versions use the same extensive set of online tutorial-style homework assignments. Engineering design activities require application of the physical science content. Optional Teaching and Learning activities help students make explicit connections between their own learning, the learning and teaching of children in elementary school, and the NGSS. With the modularity of the materials and their availability in two different formats, instructors are able to tailor their implementations in many different ways. Most instructors cover four or five of the nine units in a single term.

A large group of faculty (about 45) is currently using the *Next Gen PET* materials and participating in a collaborative *faculty online learning community* (FOLC). Research on educational transformation and the dissemination of research-based instructional strategies suggests the need to support faculty in adopting new instructional practices and materials [7-8]. In response, the *Next Gen PET FOLC* was designed to support faculty teaching with *Next Gen PET* and provide opportunities for them to learn from and support each other [9]. The community is structured in four clusters, each led by two to three experienced faculty, called cluster leaders. The project team met with cluster leaders regularly during spring 17. The remaining faculty joined the community starting in 2017-2018, met regularly via videoconference, and communicated and collaborated using online tools.

To evaluate the impacts of the materials on students' content knowledge, we developed a multiple-choice conceptual assessment that was administered pre- and post-instruction. This paper focuses on the development and field-testing of this assessment. Practical aspects of designing an assessment for a modular set of materials and of finding a common and easy administration method in a large collaboration will conclude the paper. Several authors of this paper (FG, SR, and EP) were involved in developing *Next Gen PET*. They, along with PE, are part of the leadership team for the *NGP FOLC* project, and SS is part of the external evaluation team for the project.

II. ASSESSMENT DEVELOPMENT

A. Development

There are many well-known conceptual assessments for physics topics [10]. However, no existing instrument covered the range of topics included in the NGSS-aligned *Next Gen PET* curriculum, which goes beyond mechanics to

include magnetism and static electricity; waves, sound, and light; and physical science topics such as physical changes and chemical reactions. As a result, we sought to develop a new assessment that was appropriate to the content in *Next Gen PET*. In addition to this desire for content validity, several other criteria guided development of the assessment. Horizon Research, Inc. (HRI, the project's external evaluator) and project leaders jointly generated the following design criteria for the student content assessment:

1. provides information about each of the *Next Gen PET* units;
2. can be administered in one class period;
3. consists of multiple-choice questions (this criterion was largely a consequence of the first two);
4. does not use representations that are unique to *Next Gen PET* to avoid inaccurate assessment of students' knowledge before the course.

Rather than start from scratch, a multiple-choice assessment was assembled from existing, validated items from multiple sources. Items are aligned to the *Next Gen PET* curriculum content but not to the style or representations used in the curriculum. In developing this assessment, HRI, together with the curriculum developers:

- identified the learning goals for each *Next Gen PET* unit
- collected existing, validated items from multiple sources (including Horizon's own assessments, MOSART, AAAS, and Diagnoser) and associated each item with one or more learning objectives
- explicitly excluded any items with NGP-specific style or representations
- selected approximately 5 multiple-choice items per unit (and 10 items for the "double" unit on forces)

As items were collected, they were saved in a database and coded to the science ideas in the curriculum. A handful of new items were constructed to ensure adequate idea coverage. HRI then nominated roughly twice as many items as were needed for the project leaders to review. Project leaders suggested edits to some questions and rejected others, but most were kept in their original form. Using the remaining items, HRI assembled unit-level collections of five or six items each (with the exception of the double force unit, which had 10-12 items). This resulted in a 56-item pilot assessment organized into nine sub-tests for the different content topics.

B. Pilot administration

The pilot assessment was administered by the FOLC leaders before and after their courses in spring 2017. For most of the cluster leaders, this was their first semester teaching with *Next Gen PET*. They included 4-6 units in their courses, selecting topics that aligned with their course content. The project team originally considered having faculty give the entire 56-item assessment in class. However, the cluster leaders raised a number of concerns with this approach; specifically, the amount of class time required and the potential for discouraging students by answering questions for which they had no instruction. Consequently, project leaders decided that FOLC leaders should pilot only those assessment sub-tests corresponding to the units they included in their courses. They could use their discretion whether to utilize class time or make it an outside class assignment. After surveying the FOLC leaders about their unit choices, the project assembled a customized assessment for each leader. This version was administered on bubble sheets and processed by each leader at their own institution.

C. Pilot results

Five FOLC leaders and two project team members (SR and FG) piloted the assessment in their spring 2017 classes (approximately 275 students), administering the same items during the first and last weeks of their course. The results for this pilot assessment are shown in Table 1 and Figure 1. Matched pre- and post- data are included in the *Next Gen PET* content areas of magnetism (Unit M), static electricity (Unit SE), energy (Units EM and PEF), forces (Unit IF), waves and sound (Unit WS), and light (Unit L).

Table 1: Student Performance Data

	Spring 2017							Fall 2017						
		Pre		Post		Effect	Normalized		Pre		Post		Effect	Normalized
	N	Mean	SD	Mean	SD	Size	Gain	N	Mean	SD	Mean	SD	Size	Gain
Unit M	117	75.2	19.4	94.0	9.7	1.23	0.76	303	61.7	19.9	74.6	19.2	0.69*	0.34
Unit SE	136	54.3	20.8	70.7	20.9	0.79	0.36	344	54.1	19.2	65.7	20.2	0.57*	0.25
Unit EM	267	59.2	23.4	71.4	22.1	0.53	0.30	414	49.1	24.5	62.5	23.7	0.54*	0.26
Unit PEF	212	47.7	23.7	58.4	27.2	0.42	0.20	274	49.4	18.5	65.1	21.2	0.67*	0.31
Unit IF	267	57.3	17.7	72.9	21.2	0.80	0.36	414	51.2	15.7	65.4	19.0	0.72*	0.29
Unit WS	191	63.7	26.6	75.8	23.6	0.48	0.33	270	50.1	23.5	58.0	27.2	0.31*	0.16
Unit L	185	50.0	22.5	78.8	21.9	1.30	0.58	294	41.0	20.1	64.1	23.5	0.92*	0.39

* There was a statistically significant difference between pre- and post-test scores (HLM; $p < 0.05$).

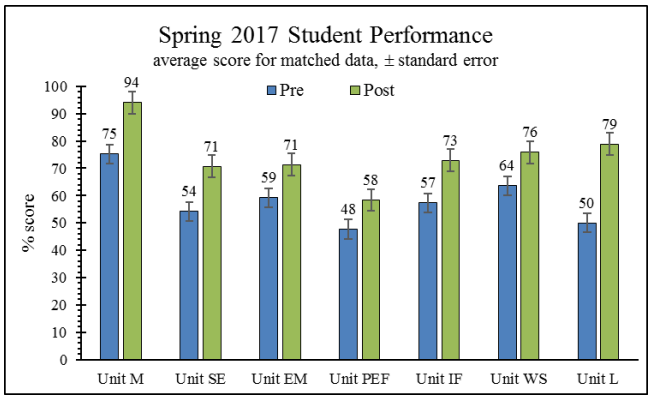

Figure 1: Spring 2017 Student Performance

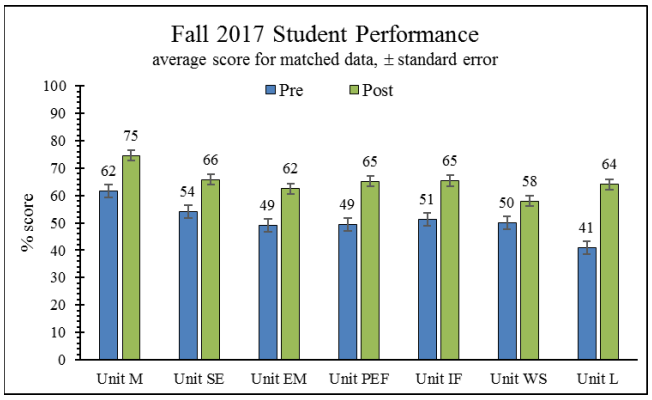

Figure 2: Fall 2017 Student Performance

The evaluation of these results centered on the assessment's ability to detect student learning gains over the course of a term, not on making claims about student performance. Evaluation of the reliability and validity of the assessment began and is still ongoing.

D. Revisions

HRI conducted classical and IRT analysis (specifically, Rasch analysis) on this pilot data, which were used to fine-tune the assessment by examining the Rasch difficulty parameter values and point biserial values. Items that were too easy and/or did not correlate well with the rest of the items on the test were removed. Four items (one each from the energy, force, light, and physical changes areas) were removed and one magnetism item was replaced. The final assessment consists of 52 items. Using Rasch analysis results, HRI generated a score look-up table to convert raw scores to Rasch scores for analysis purposes.

E. Second administration

Following pilot testing and revision, the final version of the assessment was administered by members of the FOLC who taught with *Next Gen PET* in fall 2017. After surveying the FOLC members about which units they planned to include in their courses, the project assembled a customized assessment for each member, consisting of the corresponding sub-tests. The assessment was administered outside of class for a participation grade during the first and last week of the course via one of the following formats: bubble sheets, Qualtrics, institutional learning management system (LMS), or Google Form.

F. Results

The analysis here includes only the students of 17 instructors whose fall 2017 data was available at this time. Before analyzing the data, HRI conducted Rasch [11] analyses to establish difficulty parameters, which were then used to create a Rasch score for each student's pretest and posttest using the lookup table mentioned above. The scores were transformed to a 100-point scale for analysis and reporting.

Because instructors administered only the items aligned to the units they taught, HRI analyzed students' scores for each of the nine sub-tests. The analysis involved comparisons of conceptually related outcomes (i.e., scores on tests that address related content). For instance, there are two units on energy; though these have different foci, the content is closely related. Therefore, the comparisons of pre- and post-test scores for these two units are not strictly independent. Statistically, they share a common pool of error. For this reason, the False Discovery Rate (FDR) method [12] was used to adjust the alpha level required for statistical significance and maintain an overall Type I error rate of five percent. Statistical tests on unit sub-test scores within a module were adjusted as a set using this procedure, except Interactions and Forces, which was treated as its own set with no adjustments.

The student scores included in the analysis are not independent of one another but rather are grouped in classes. Therefore, their scores share a source of variance (e.g., all students in the same class had the same instructor), making them, like the students, not completely independent. Hierarchical linear modeling (HLM) takes into account this shared variation and is therefore more statistically powerful than analyses that do not group student scores within instructors. HLM was used to examine differences between pre- and post-course scores on each sub-test, nesting student scores within instructors. The analysis found a statistically significant difference between the pre- and post-test score for each sub-test (HLM; $p < 0.05$); Figure 2 and Table 1 summarize these results. Most of the effect sizes are moderate or moderate to large, suggesting the gains were meaningful in addition to being statistically significant.

The effect sizes in Table 1 suggest apparent differences among unit sub-tests in terms of student learning gains. However, it is important to bear in mind that each sub-test typically included 5 or 6 items, so the measures are not very precise, making claims about differences among units unwarranted. What seems clear, however, is that despite the

encouraging effect sizes indicating substantial gains, students still had room for growth at the end of their course.

III. PRACTICAL MATTERS

The *NGP FOLC* provides an important opportunity to gather student impact data from a large number of courses at different institutions. Effectively capitalizing on this opportunity requires an approach that is acceptable to the participating faculty, while also satisfying project requirements for quality and standardized administration. Additionally, the modular nature of *Next Gen PET* means that different instructors will teach different combinations of units, complicating administration and analysis of the assessment. The project team and FOLC leaders discussed these issues and developed an approach that best satisfied the sometimes competing requirements of the project and individual faculty. This approach is described below.

In fall 2017, we adopted an online administration procedure that incorporates best practices for online administration based on the research literature [13]. Each term instructors complete a Qualtrics survey, which provides the start and end dates of the term and the units to be covered. Using this information, the project team (PE) provides each instructor with a customized Google Form version of the assessment. Instructions lead an instructor through setting up the Google Form locally and sending a link to their students. All data is recorded on the instructor's Google Drive. This ensures that the data remains in the control of each instructor and thus, avoids a more complicated IRB approval process. At the end of the term, instructors are guided through the Data Explorer upload process.

PhysPort's Data Explorer [14] is an online tool for scoring, analyzing, and interpreting results. The PhysPort Data Explorer team has developed queries and resources specifically for the *NGP FOLC* project. Instructors retain control over their data and can utilize it in future, action research projects. As the NGP assessment database grows, instructors will eventually be able to compare their results with national data to help them understand their results. This online tool streamlines management of data from nearly 50 different instructors, tracks course details (format, size, etc.), and allows HRI and project staff to query the project data set.

IV. CONCLUSIONS

Assessing student impacts presents additional challenges when the content included in a curriculum does not align with existing assessments. This is the case with *Next Gen PET*; as a result, we developed, piloted, and revised a tailored assessment consisting of existing items. This approach represents a middle ground between developing an assessment from scratch and using an existing assessment. The modular nature of *Next Gen PET* provides additional challenges to administration and analysis. Instructors administered a "customized" version of the assessment including only sub-tests for the units they covered, and the results were analyzed at the unit level.

ACKNOWLEDGEMENTS

Thanks to all the *Next Gen PET Faculty Online Learning Community* members who contributed data to this analysis. Thanks to the PhysPort Data Explorer team for their contributions. This work is supported by funding from NSF-1626496.

[1] NGSS Lead States. *Next Generation Science Standards: For States, By States*. (National Academies Press, Washington, DC, 2013); National Research Council. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. (National Academies Press Washington, DC, 2012).

[2] http://ngss.nsta.org/About.aspx. Retrieved 7/4/2018.

[3] National Research Council. *Taking Science to School: Learning and Teaching Science in Grades K-8*. (National Academies Press, Washington, DC, 2007).

[4] F. Goldberg, S. Robinson, and V. Otero, *Physics and Everyday Thinking*. (Activate Learning, Greenwich CT, 2007).

[5] F. Goldberg, V. Otero, and S. Robinson, Am. J. Phys. 78, 1265 (2010).

[6] F. Goldberg, E. Price, S. Robinson, D. Boyd-Harlow, and M. McKean. Phys. Rev. ST Phys. Educ. Res. 8, 010121 (2012).

[7] C. Henderson, A. Beach, and N. Finkelstein, J. Res. Sci. Teach. 48, 952-984 (2011).

[8] C. Henderson, M. Dancy, and M. Niewiadomska-Bugaj, Phys. Rev. ST Phys. Educ. Res. 8, p. 020104 (2012).

[9] http://ngpfolc.com

[10] A. Madsen, S. McKagan, and E. Sayre. AJP 85, 245 (2017)

[11] Boone, W. J. CBE—Life Sciences Education, 15(4), rm4. (2016).

[12] Benjamini, Y. & Hochberg, Y. *Journal of the Royal Statistical Society,* B, 57, 289–300. (1995).

[13] Nissen, J. M., Jariwala, M., Close, E. W., & Dusen, B. V. *International Journal of STEM Education*, *5*(1). (2018).

[14] https://www.physport.org/DataExplorer/Preview.cfm