Balancing Development and Measurement Needs in an Evaluation of a Program Under Development

George M. Harrison, Lisa M. Vallin, Joanna Philippoff, Paul R. Brandon, & Kanesa Seraphin
Curriculum Research & Development Group
University of Hawai‘i at Mānoa

April 2014, AERA
Purpose
To describe what we did and what we learned in:

- measuring outcomes in a program under development
- meeting the request to include a comparison group
- dealing with the nested-structure of the data
Background

- Responsive evaluation
  - rather than preordinate evaluation (Stake, 1973) is appropriate for evaluating a program in development.

- Involving stakeholders
  - in the evaluation builds credibility (Yarbrough, Shulha, Hopson, & Caruthers, 2011).

- Balancing rigor and feasibility
  - is important in making decisions about measuring outcomes (Braverman, 2013; Brandon, Lawton, Harrison, 2014).

- Multilevel modeling
  - is appropriate for examining student-level outcomes with teacher-level moderating variables such as implementation and prior knowledge (Zvoch, 2012).
Investigating the Effects of a Science Professional Development (PD) Intervention

Teaching Science as Inquiry: Aquatic

2010-2012
O’ahu

2011-2012
Maui
Big Island

2012-2013
O’ahu
Kaua’i
Pre- & Post-PD Instruments

Multiple PD workshops throughout the school year

August
Pre

Student Science Questionnaire
Teacher assessments and questionnaires

Post

Teacher Interviews

May
Pre-Post Quasi-Experimental Design

**Comparison group**
- Early versions of PD workshops
  - Maui & Big Island
- August, 2011 to May, 2012

**Treatment group**
- Refined versions of PD workshops
  - O’ahu & Kaua’i
- August, 2012 to May, 2013

Changes in students’ understanding of the nature of science (NOS)

August, 2011

August, 2012

May, 2012

May, 2013

Pre

Post

August, 2011

August, 2012

May, 2012

May, 2013
Measuring Students’ Understanding of NOS: Defining the Content Domain

- Discussions with the PD developers
- TSI theoretical framework (Pottenger; Tice & Duncan, 2009)
- Previous studies and frameworks (Ayala, 2004; Brandon, et al. 2007; National Research Council, 2012)
Measuring Students’ Understanding of NOS: Item Development

- Searched for existing instruments

- Adopted Ayala’s (2005) and Brandon et al.’s (2007) existing scale

- Drafted new, best answer, MC items

- Discussions with the PD developers

- Existing Likert-scale items (Ayala, 2005)

- New MC items
Measuring Students’ Understanding of the NOS: Item Editing

- Review by content experts
- Discussions w/ the PD developers
- Empirical data from early pilot testing with University Laboratory School students
- Focus-group interviews with the students
- Empirical IRT data from a second pilot testing (TSI students in 2011-2012)

Likert-scale items
Best-answer MC items
Measuring Students’ Understanding of the NOS: Item Editing

Edits by and with the PD developers based on empirical data and project refinement

Empirical IRT data from a third group of students (TSI students in 2012-2013)

Equating later form with earlier-year form, to compare the groups in their pre-to-post change

- Likert-scale items
- Best-answer MC items
The Model

Changes in students’ NOS

Implementation variables (adherence and exposure)

Prior pedagogical content knowledge (PCK)

Prior understanding of inquiry

High or middle school

Year in program development
## Multilevel Model Results

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Outcome variable (NOS scores)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.93***</td>
</tr>
<tr>
<td>Pre-post change&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.19***</td>
</tr>
<tr>
<td>(Year in the program)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>-0.24</td>
</tr>
<tr>
<td>(Year in the program)&lt;sup&gt;b&lt;/sup&gt; *(Pre-post change)</td>
<td>-0.01</td>
</tr>
<tr>
<td>School level&lt;sup&gt;c&lt;/sup&gt;</td>
<td>-0.17</td>
</tr>
<tr>
<td>(School level)&lt;sup&gt;c&lt;/sup&gt; *(Pre-post change)</td>
<td>-0.16*</td>
</tr>
<tr>
<td>(Pre PCK)&lt;sup&gt;d&lt;/sup&gt; *(Pre-post change)</td>
<td>-0.01</td>
</tr>
<tr>
<td>(Pre ITA)&lt;sup&gt;e&lt;/sup&gt; *(Pre-post change)</td>
<td>-0.01</td>
</tr>
<tr>
<td>(Adherence)&lt;sup&gt;d&lt;/sup&gt; *(Pre-post change)</td>
<td>0.06</td>
</tr>
<tr>
<td>(Exposure)&lt;sup&gt;d&lt;/sup&gt; *(Pre-post change)</td>
<td>-0.07</td>
</tr>
</tbody>
</table>

Note. NOS scores are in logits (partial credit IRT model) after equating the earlier and later tests. The model is a multilevel mixed model, with Pre-post change as Level 1, Student as Level 2 (random intercepts and slopes), and Teacher (random intercepts) as Level 3. Main effects were omitted from table for brevity.

<sup>a</sup>Pre-post change was coded 0 for pre and 1 for post.  
<sup>b</sup>Year in the program was coded 0 for SY 2011-2012 and 1 for 2012-2013.  
<sup>c</sup>School level was coded 0 for middle and 1 for high school.  
<sup>d</sup>Each of these was standardized $(M = 0, SD = 1)$.  
<sup>e</sup>ITA was on a logit scale $(M = 0)$.

* = $p < .05$; *** = $p < .001$. 
We addressed the following needs:

- The program’s development
  - The very process of collaboration in developing the instrument helped the program refine its intended outcome
- The instrument’s development
  - We dealt with threats to validity
    - Content validity: included experts, PD developers, students, and the program’s theory
    - Construct validity: included both self-report and direct measure
- The inclusion of a comparison group
  - We equated instruments under different stages of development, permitting group comparison
- Nested structure of the data
  - teacher-level prior knowledge and fidelity of implementation
What we learned

- measuring outcomes in a program under development
  The outcomes-measure development
  a) helped the evaluators to better understand the PD and provide feedback
  b) prompted the PD developers to carefully conceptualize their intended outcome

- meeting the requirement to include a comparison group
  We achieved this, but would not recommend our approach because of weak program differentiation. Also, there were covariates (conflated with group) we could not measure.

- dealing with the nested-structure of the data
  We dealt with this, but found no effect. More time points would have strengthened the results and permitted the inclusion of mediating variables (such as teacher-level proximal outcomes).
Limitations

- We were restricted in the number of covariates we could use; could not account for variables that could have explained cohort membership.
  - Because background data-collection instruments were also under development
- Could not collect student-level covariates, such as English language arts scores, that could rule out alternative explanations for growth.
Acknowledgement

- The research reported here was supported by the Institute for Educational Sciences, US Department of Education, through Grant R305A10091 to the University of Hawaii (UH) at Manoa.