Validating an Assessment through Rasch techniques:

The process of measurement, feedback, reflection and change

Kelly D. Bradley and Shannon O. Sampson

University of Kentucky
Abstract
The Hands-On Virtual (HOV) Workshops implement a content driven, inquiry-based physics curriculum for upper elementary and middle school teachers, delivered in a distance-learning format. As a facet of the workshops, pre- and post- course multiple-choice assessments were constructed and administered to 43 teacher participants with the aim of measuring growth in teacher content knowledge. This study applied the Rasch model to determine: if all items functioned as expected, how well the tests were targeted to the ability of the examinees, how well the items were distributed along the continuum of the variable and how the examinees used the distracters. Recommendations are provided to revise the assessment for future use. Results of the study provide a methodology for validating assessments.
Validating an Assessment through Rasch techniques:

The process of measurement, feedback, reflection and change

Given the gap in student achievement in math and science in Appalachia, there is a need to provide connections among teachers, students and higher education resources. Distance learning teacher training provides a medium for making these connections. Little research has been conducted in regard to the efficacy of implementing hands-on, inquiry-based science instruction using a distance learning format. While there is a need to assess growth in teachers’ scientific content knowledge, both in the distance learning format and other contexts, few well-designed assessments have been constructed for this purpose. One task of the National Science Foundation (NSF) funded research project, *Assessing How Distance Learning for Teachers Can Enable Inquiry Science in Rural Classrooms*¹, is to restructure teacher assessments to support the investigation of whether delivery of scientific content through a distance-learning format improves middle school teacher science knowledge.

The NSF project is working in conjunction with Appalachian middle schools in Southeastern Ohio, West Virginia, Eastern Kentucky, Virginia and Eastern Tennessee. Dr. Joe Straley and Sally Shafer at the University of Kentucky via a FIPSE grant developed the Hands on Virtual Physics (HOV) course. HOV science courses are delivered in a hands-on distance learning format to elementary and middle school teachers. Currently, there are three courses offered in connection with the HOV curriculum: Light, Electricity & Magnetism, Temperature and Heat. Presently, the focus is on the Temperature and Heat course, which includes 10 units of study that explore thermal expansion, liquid crystals and thermal conduction (see [http://www.pa.uky.edu/~sciworks/heat/preview/frame.htm](http://www.pa.uky.edu/~sciworks/heat/preview/frame.htm) for more details).

---
¹ NSF 04-553; REC Educational Research Initiative. *Assessing How Distance Learning for Teachers Can Enable Inquiry Science in Rural Classrooms.* ($1,998,833); 1/1/05 – 12/30/09.
The task of constructing assessments to responsibly measure teacher and student learning gains through their connection with the Hands on Virtual Physics workshops takes a team approach of individuals with expertise in education, assessment, measurement and the specific content. Given the limited number of well-designed assessments connected to the distance learning format, instruments are needed to assess growth in teachers’ scientific content knowledge and pedagogical skills. An integral component of this paper is to describe the process of restructure an existing exam so that it is a valid assessment supporting the study of the teacher science knowledge outcome.

Objectives

The Rasch model described below, along with a descriptive summary, was applied to respond to the following questions.

1. Are items functioning as expected?
2. How well is the test targeted to the ability of the examinees?
3. How well are the items distributed along the continuum of the variable?
4. How did the examinees respond to each of the items?

Method

Data Source

The Hands-On Virtual (HOV) Physics workshops, developed under FIPSE support, constructed pre- and post- course multiple-choice and open-response assessments that were administered to 43 upper-elementary and middle school teachers. These assessments were implemented in an effort to measure the growth in teacher content knowledge connected to the training. The focus of this study is a restructuring of the multiple-choice assessment, which in its original form consisted of 23 items with five choices, one of which was the correct response.
Specifically, attention is given to teacher data from the post-course assessment, to create an assessment that is calibrated with those that have experienced the training resulting in setting a realistic, but high enough standard measure.

Data Analysis

The dichotomous Rasch model was employed to analyze the post-course assessment data. Here, Rasch model calibration and measurement are based on two expectations: a more able person should have greater probability of success on items than a less able person, and any person should always be more likely to do better on an easier item than on a harder one. In contrast to Item Response Theory models, in which models are constructed to fit the given data, the Rasch model takes a prescriptive approach, asking the underlying question of how well the data fit the model assumptions. Analyses were conducted within WINSTEPS. The model being employed is a stochastic parameterization of Guttman data. Specifically, for the multiple-choice assessment, the measurement model is

\[
\log \left( \frac{P_{nix}}{1 - P_{nix}} \right) = B_n - D_i
\]

where \( P_{nix} \) is the probability of a successful response being produced by person \( n \) to item \( i \), \( B_n \) is the ability of the person \( n \) and \( D_i \) is the difficulty of the item.

Results and Discussion

Issues for multiple-choice items include selection of distracters, the unequal distribution of questions as compared to examinee distribution, effectiveness of item wording and practice effect. The addition of ‘more difficult’ multiple-choice items is recommended. The processes for reconstruction of the assessments, based on Rasch analyses results, are reviewed.

Measurement and Feedback

In an effort to construct a reliable and valid assessment, the focus was placed on measurement. Bond and Fox (2001) state:
“Operationalizing and then measuring variables are two of the necessary first steps in the empirical research process. Statistical analysis, as a tool for investigating relations among the measures, then follows. Thus, the interpretation of analyses can only be as good as the quality of the measures” (p. xvi).

The assessment under review was originally constructed with the intention to measure content growth attributed to the HOV training for teachers on the temperature and heat unit. Thus, the items are theoretically argued to be unidimensional. To assess the validity of the instrument, research questions were answered for each item on the assessment.

The process of measurement and feedback began with investigating the fit of the items to the model, answering the question: Are all items functioning as expected? Table 1 displays the item measures and fit statistics.

Table 1. Item Measures and Fit Statistics
Validating an Assessment

Viewing the item measures and fit statistics (See Table 1) provides a response to the first research question, which specifically addresses if the items are fitting the expectations of the Rasch model. ZSTD fit statistics should fall between –2 and 2 and point measure correlations should all be positive, supporting the assumption of unidimensionality. Item 23 has large infit and outfit ZSTD scores and is a negative correlate, so it was highlighted for further review.

In view of how well the test targeted the ability of the examinees, the variable maps should be considered (see Figure 1). In a well-targeted test, mean item and person measures are approximately equivalent on the vertical ruler placing persons and items on the same scale. The ruler is labeled on both sides with the mean as well as one and two standard deviation indicators. The left side displays the 43 teachers completing the training, while the right side displays the 23 items on the assessment. Here, the mean person measure is above the mean item measure, so it appears the test is somewhat easy for this group. In a pre-post test application, this is not unreasonable. Given the conjecture that the training will improve participant knowledge, it is acceptable for the mean person measure, teachers’ average, to be higher than the item measure.
Because of the distribution of items and teachers on the map, attention is directed at how well the items are distributed along the continuum. Figure 2 displays the same output presented in Figure 1; however, here the gaps in the item distribution are designated.

Figure 2. Person and Item Map
In a well constructed assessment, the items spread evenly over the entire continuum of teacher ability. In this case, two gaps are present. One gap is located near one standard deviation from the mean of the items. For the eleven teachers in that gap, the measure of their ability level is estimated by items that are either well below or above their ability level. In order to generate more meaningful information about the ability of teachers falling in this range, items should be constructed that are more difficult than item 3, but easier than items 11 and 12.

The second gap exists at the top of the continuum, where a few teachers have a much higher ability than any item assessed. Again, in structuring an assessment, it is important to cover the entire person ability level to generate meaningful information about all persons. Using the Rasch output, it is suggested that a few items be written that are more difficult than item 17.
Response patterns must also be considered. Table 2 presents a sample of items and the empirical responses to these items, listed according to how poorly the item fits the assumptions of the model. The correct option is listed last and highlighted. The table includes the percentage of teachers that selected each choice and the average ‘ability’ measure of the teachers selecting that option, with the mean measure being set at 50.

As previously discussed, an underlying assumption of the Rasch model is a hierarchy of difficulty and ability. Simply stated, it is not logical for teachers to choose a distractor, a wrong choice, if they have a higher average ‘ability’ measure than teachers who selected the correct choice for that item. The answer choices not fitting that assumption are circled. The outfit mean-square (mnsq) values in the chart indicate the degree to which the responses fail to meet the assumption. Values greater than 2 are problematic in that they likely adversely affect quality of the instrument. Revisiting item 23, which was a negative correlate of the other items, distracter e of item 23 was selected by many teachers with a higher measure than those teachers selecting the correct option, such that the outfit mnsq was 6. This item was reviewed for vague wording, confusing answer options and/or a possible error in the solution key.
Table 2. Sample of Response Analysis

**Reflection and Change**

Drs. Becky McNall and Joe Straley, along with Sally Shafer offered insight into issues with the selected items and rewrote the items based upon the feedback provided. Items 23, 17, 3 and 9 received the greatest scrutiny. Changes made to items 17 and 23, the most problematic items, are outlined below.

**Reflection for Item 17**

Figure 3 displays item 17 from the original assessment. It includes annotations from the assessment committee, noting that the wording may have been confusing enough that higher ability teachers were drawn to the incorrect answer.
Old Post 17) When object A is placed in contact with object B, thermal energy flows from A to B. What does this imply?

⇒ Concept A, answer d.

This is the hardest question on the test. They have trouble choosing between answers b and and/or d. This is asking two things. We need to reword to ask only one question, perhaps by making two questions of it. Using ‘imply’ is bad form. This may benefit from visual presentation.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>a. A is more massive than B</td>
</tr>
<tr>
<td>4</td>
<td>b. A contains more energy than B.</td>
</tr>
<tr>
<td>1</td>
<td>c. A is an insulator and B is a metal</td>
</tr>
<tr>
<td>11</td>
<td>d. A has a higher temperature than B,</td>
</tr>
<tr>
<td>27</td>
<td>e. b and d are both correct</td>
</tr>
</tbody>
</table>

Item 17, presented below in Figure 4, was changed accordingly to address the above concerns.

Figure 4. New Item 17

When object A is placed in contact with object B, thermal energy flows from A to B. Which statement is always true?

a. A has both a higher temperature and more energy than B.

b. A contains more energy than B.

c. **A has a higher temperature than B.**

d. A is a better conductor than B.
Because the Rasch analysis highlighted the problematic nature of the wording, the committee was able to write a more concise question, including removing one distracter and writing out the original distracter e (now a). A follow up analysis would reveal if this is still the most difficult item on the assessment or if it was only difficult because a number of teachers found it confusing.

Reflection for Item 23

Figure 5 displays the original item 23, also including the annotations written by the assessment committee.

Figure 5: Original Item 23

Old Post 23) Imagine a “thermosicle” – it’s like a Popsicle made using water and a thermometer for a stick, so the thermometer bulb is frozen inside the ice. If it’s taken out of the freezer, placed in a dish on the counter top, and left to melt, which graph best represents how the temperature of the “thermosicle” changes over time?

⇒ Concept E, answer b

People who otherwise are scoring high are missing this one, scoring below less accomplished learners – perhaps due to over analysis or poor distracters leading them astray. Distracter e is their choice.
Recall, this item fit very poorly to the assumptions of the Rasch model, mainly because of option e distracting some high ‘ability’ teachers. The original format of Item 23 does not offer a 0 temperature point. This was included when the item was restructured. It was also recognized that ‘more able’ individuals might put more thought into what happens early in the process, leading them to misinterpret the graphs. Based upon the analysis and conversation, the item answer choices were reworked as displayed below in Figure 6.

Figure 6. New Item 23
Conclusion

The analysis proved useful in identifying problematic features of the assessment that the research team was able to incorporate into the restructuring of the teacher assessment. Results indicated that the assessment was reasonable but had room for improvement. Using the Rasch analysis results, items were reviewed and suggestions for restructuring were suggested. Examinees response patterns in connection to the distracters were given much consideration. Many distracters seemed to have presentations that were confusing to teachers, making the item less about knowledge and more about not falling for any ‘tricks’. Repeatedly, only a few distracters were selected as teachers selected answers for items, with very few items having teachers utilize all 4 distracters. Because of this, the new assessment will remove a distracter from each item, reducing the number of distracters from 4 to 3 distracters, providing 4 choices total. Additional items are being written to address the gaps discussed in the person and item continuum. Finally, to complement the multiple-choice assessment and create an authentic measure of teachers’ content knowledge, open-ended questions and performance tasks are also being constructed.

The recommendations from this study were used to restructure and improve the teacher Temperature and Heat assessment for future testing of teacher content knowledge of Temperature and Heat as related to the Hands on Virtual Physics workshop. This effort will provide a more accurate picture of teacher knowledge in future administrations, creating an assessment that can be used to measure growth in content knowledge. The study also outlines Rasch measurement for use with assessment data, providing a methodology for validating assessments. It is a systematic and straightforward process of evaluating the quality of
assessment instruments, one that the general practitioner could use to ensure valid and comprehensive pictures of targeted outcomes.

References

