Applying the Rasch Model to Measure Change in Student Performance over Time

Jessica D. Cunningham; Western Carolina University

Kelly D. Bradley¹; University of Kentucky

¹ Please use Kelly D. Bradley as author of contact. 131 Taylor Education Building, Lexington, KY 40506-0001, kdbtrad2@uky.edu or (859)257-4923
Applying the Rasch Model to Measure Change in Student Performance over Time

Abstract

*Newton’s Universe* (NU), an innovative teacher training program, strives to obtain measures from rural, middle school science teachers and their students to determine the impact of its distance learning course on understanding of temperature. Although limited research exists examining the usefulness of the Rasch model in measuring change over time, this study applied the method of racking and stacking the NU data to examine change in student performance over time. Locally, researchers involved in the NU course development can use these findings to further improve the course and the student assessment instrument. Researchers across various fields can use the methods shown in this study to improve assessment instruments, course development, and measures of change over time.

**DESCRIPTORS:** Rasch Model, Racking and Stacking, Measuring Change in Student Performance, Newton’s Universe teacher training
Applying the Rasch Model to Measure Change in Student Performance over Time

Challenges regarding student achievement in math and science education remain an issue for rural educators in general and particularly in the rural regions of Appalachia. Newton’s Universe (NU)\(^2\) is an inquiry-based teacher training program striving to obtain measures from rural, middle school science teachers and their students to determine the impact of its distance learning course on understanding of temperature. The goals were to enhance the conceptual base of science educators in state and national content standards in physical science, introducing and developing inquiry strategies, integrating technology into teaching and learning, and ultimately improving student learning. Teacher knowledge is instrumental in impacting student achievement (Darling-Hammond, 1999; NCLB, 2002; Rowan, Correnti, & Miller, 2002), but rarely discussed as impacts of teacher training, are measures of change in student understanding. This is likely due to the many potential issues associated with measuring change over time in psychological constructs such as isolating intervention effects. Although typical for a professional development program to administer an assessment to measure its impact on teachers, essential to the evaluation of professional development in education is evidence of student learning associated with an intervention (Guskey, 2000), such as student baseline and post instruction measures.

NU provided participating teachers a distance learning course during the summers of 2006 through 2008 with an emphasis on inquiry-based learning of temperature topics. In addition to obtaining teacher baseline and post instruction measures from control and experimental groups, teachers administered baseline and post instruction assessments to their middle school science students. Measures for teacher (Bradley & Sampson, 2006) and student understanding of

\(^2\) Newton’s Universe funded by the National Science Foundation Grant No. 0437768 (J. Osborn, P.I.). Further information can be found at the project website: http://www.uky.edu/NewtonsUniverse.
topics covered in the course were taken using a quality assessment developed by the NU research team. This study illustrates the process of applying the Rasch model to measure change in student understanding of temperature over the time and ways these results can be used to inform course developers as well as teachers participating in the project.

**Purpose of Study**

Limited research exists featuring the Rasch model and its usefulness in depicting change in measures over time. The purpose of this study is to apply the method of racking and stacking the NU student data within the Rasch model to address the following research questions:

1. Are items on the baseline and post instruction science teacher assessments functioning as expected within the racked analysis? In other words, are post instruction item estimates less than corresponding baseline item estimates?

2. How well are the items distributed along the continuum of the “understanding of temperature” variable in the racked analysis? How does this distribution change from baseline to post instruction assessment?

3. Are persons functioning as expected from baseline to post instruction assessment? In other words, are student estimates increasing from baseline to post instruction assessment?

4. What are the plausible implications of these results for NU and others?

The NU student data were examined under the assumptions of the Rasch model, specifically assessed for unidimensionality. In the racked analysis, the purpose was to examine the change in items over time while the focus of the stacked analysis was to examine change in student performance over time (Wright, 2003). Items and persons were placed along a continuum where examinees are measured based on their performance in comparison to where the items are measured based on their difficulty. A wider spread of items allows for the measurement of a
larger range of person abilities. Investigating item difficulties on the same ruler within the racked analysis illustrates change in items over time. The expectation would be that item estimates decreased over time. In other words, item difficulties should be higher for the baseline assessment compared to the post assessment. Placing baseline and post assessment student performance estimates along the same continuum allows the examination of whether each student measure increased from baseline to post assessment. Transforming measures onto the same ruler enables meaningful implications of these results for NU and others.

**Literature Review**

The launching of Sputnik I by the Soviet Union in 1957 spawned the appropriation of federal funds to place an emphasis on science education, deemed imperative for our national security. Building on the foundation of inquiry-based learning (Hunkins, 1972), NU is a distance learning program focused on the inquiry-based learning of physical science for middle school teachers with an interest in discovering its impact on middle school student learning. While the impacts on teachers are typically measured, it is a rare opportunity to also measure transferability of teacher knowledge to students as a result of an intervention. Equally important to developing assessments to measure growth in understanding is responsibly reporting the results to accurately reflect the measures in a meaningful way for not only the course developers but also its participants.

Although more than one method of analysis exists to determine growth in understanding over time, item response theory (IRT) has noteworthy additions to the traditional classical test theory (CTT) approach (Embretson & Reise, 2000). This study features the application of the Rasch model equivalent to the one parameter, specifically item difficulty, IRT model. The following provides a review of literature to build a foundation for applying the Rasch model to
Rasch Model to Measure Change

measure change in understanding over time. The review begins with outlining important additions of the Rasch model compared to CTT followed by ways to determine fit of the data to the Rasch model. After theoretical and empirical fit of the data to the model have been established, a conceptual foundation for racking and stacking the data is presented.

From a theoretical standpoint, the Rasch model provides important additions to the measurement of change to CTT. The Rasch model consists of "model-based measurement in which trait level estimates depend on both the persons' responses and on the properties of the items that were administered" (Embretson & Reise, 2000, p. 13). The Rasch model theorizes an additive linear structure underlies the observed data, that both participants and items can be arrayed on a continuum, and that the items are equivalent in item discrimination (Bond & Fox, 2001). Rasch measurement produces a standard error (SE) for each person and item, specifying the range within which each person’s ‘true’ performance and each item’s ‘true’ difficulty fall when a person or item fits the model and the corresponding SE is small. Individual errors can then be used to produce a more accurate average error variance for the sample (Wright & Masters, 1982). The property of invariance – the expectation that the measures remain the same regardless of which items are used as well as independent of the sample used – is an essential component of Rasch measurement allowing interpretations over useful range of time and place (Wright & Masters, 1982).

Prefacing any analysis, the data must uphold the required assumptions of the model. In this case, the underlying assumption of the Rasch model is unidimensionality, referring to measuring only one attribute at a time (Bond & Fox, 2001). Theoretically, the NU student assessment was designed to measure a unidimensional construct – specifically understanding of temperature. Although content experts argue that unidimensionality holds from a theoretical
standpoint, fit statistics along with a principal component analysis can determine if the assumption indeed holds empirically (Bond & Fox, 2001). Wright and Stone (1994b) recommend the guideline of 0.7-1.3 as an acceptable range for infit and outfit mean square statistics for a multiple choice assessment. In other words, fit statistics outside a literature-based established cutoff indicate a violation of expected response patterns central to Rasch measurement (Smith, 2004): that a person with a greater ability has a higher probability of correctly answering any item and that an item of lesser difficulty has a greater probability of being answered correctly (Wright and Masters, 1982). Assuming the data fit the model, interpretations of results may proceed.

The assessment was examined using the Rasch model similar to Bradley and Sampson (2006), and appropriate changes were made to improve the quality of the instrument. Therefore, this study used the restructured assessment and focused on applying the Rasch model to examine change in measures over time. Wright (2003) posed a method of racking and stacking the data within the Rasch model to determine change in student performance (stacking) as well as change in item difficulties (racking) over time. Although the usefulness of racking and stacking the data has been posited by Wright (2003), limited research exists applying the Rasch model for this purpose.

In order to employ the Rasch model to depict change in measures over time, the data must be restructured according to Wright (2003) by racking and stacking the data. Stacking the data is done by appending the person measures for the post instruction onto the baseline measures resulting in twice as many persons being measured (see Figure 1).
Conceptually, this transforms both the baseline and post instruction assessment measures into measures on the same ruler to make interpretations about the effect of the instruction on student understanding. The assumption is that the students have changed over time as a result of instruction prior to the post assessment. Similarly, racking the data can be done by attaching the post instruction item responses onto the baseline item responses resulting in each person taking twice as many items on the assessment (see Figure 1). This conceptual approach illustrates item difficulty estimates from baseline and post instruction assessment on the same ruler to make conclusions about the effect of the intervention on the change in items. Racking the data allows the inspection of change in the items rather than students by the end of instruction. In summary, "stacking the data, we see who has changed" while "racking the data, we see what has changed" (Wright, 2003, p. 906). This study applies the Rasch model to the restructured NU data to illustrate its appropriateness in examining change in measures over time. Followed by assessing the data for unidimensionality, results are presented and interpretations are discussed to provide information to NU and its participants.
Methodology

Although analysis of variance is often the method of choice in educational research to depict change over time, others exist with accurate results and potentially more meaningful interpretations for interested parties. The importance of choosing accurate and meaningful analyses requires attention in order to evaluate impacts of interventions such as a teacher training program. The purpose of this study was to illustrate the usefulness of the Rasch model in depicting change in measures over time and implications of these results.

Sample

The selection of teachers is first detailed because the student sample is dependent on the participation of their teachers in the NU program. The target population for the NU project was middle school science teachers from the sampling frame of all public high schools in rural regions of Appalachia, excluding teachers who had participated in any Appalachian Mathematics and Science Partnership (AMSP) training workshops of physical science. As a participant, teachers were asked to complete surveys and assessments on a voluntary and confidential basis in exchange for stipends and classroom materials. In an effort to isolate the NU course impacts, control teachers were selected under the same criteria as participants and offered stipends with the opportunity to enter the project as an experimental teacher the following summer. Descriptive statistics of demographic data for the teachers have been produced (see Table 20).
### Table 20

**Descriptive statistics for teacher demographics**

<table>
<thead>
<tr>
<th>Demographic Variable</th>
<th>Number of Teachers (n = 47)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grade Level Currently Teaching</strong></td>
<td></td>
</tr>
<tr>
<td>6th only</td>
<td>12</td>
</tr>
<tr>
<td>7th only</td>
<td>8</td>
</tr>
<tr>
<td>8th only</td>
<td>11</td>
</tr>
<tr>
<td>6th and 7th</td>
<td>10</td>
</tr>
<tr>
<td>7th and 8th</td>
<td>1</td>
</tr>
<tr>
<td>6th, 7th, and 8th</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
</tr>
<tr>
<td><strong>Subjects Teaching</strong></td>
<td></td>
</tr>
<tr>
<td>Science only</td>
<td>31</td>
</tr>
<tr>
<td>Science and other</td>
<td>15</td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
</tr>
<tr>
<td><strong>Highest Level of Formal Education Completed</strong></td>
<td></td>
</tr>
<tr>
<td>B.A.</td>
<td>4</td>
</tr>
<tr>
<td>B.S.</td>
<td>10</td>
</tr>
<tr>
<td>Master’s</td>
<td>21</td>
</tr>
<tr>
<td>Ed.D./Ph.D.</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
</tr>
<tr>
<td><strong>Science Certifications Held</strong></td>
<td></td>
</tr>
<tr>
<td>Physics</td>
<td>2</td>
</tr>
<tr>
<td>Chemistry</td>
<td>5</td>
</tr>
<tr>
<td>Physical science</td>
<td>3</td>
</tr>
<tr>
<td>Biology/life science</td>
<td>16</td>
</tr>
<tr>
<td>Earth/space science</td>
<td>4</td>
</tr>
<tr>
<td>General science</td>
<td>17</td>
</tr>
<tr>
<td>No certification</td>
<td>11</td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
<tr>
<td><strong>Years Teaching Science</strong></td>
<td></td>
</tr>
<tr>
<td>0-2</td>
<td>11</td>
</tr>
<tr>
<td>3-5</td>
<td>10</td>
</tr>
<tr>
<td>6-10</td>
<td>10</td>
</tr>
<tr>
<td>11-15</td>
<td>6</td>
</tr>
<tr>
<td>16-20</td>
<td>8</td>
</tr>
<tr>
<td>Over 20</td>
<td>2</td>
</tr>
</tbody>
</table>
The student sample was a relatively homogeneous group with similar backgrounds from rural Appalachia. These students are in middle school science classes (6th, 7th, and 8th grades) that are taught temperature topics by control and experimental NU teachers. Each student was administered a subset of an assessment based on the temperature topics taught by their teacher. The number of students administered either the entire assessment or a subset of the assessment has been summarized (see Table 1).

### Table 1

*Number of teachers by cohort across years*

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Teachers</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Students</td>
<td>ABCDE/All</td>
<td>ABCDE/All</td>
<td>ABCDE/All</td>
</tr>
<tr>
<td></td>
<td>Base</td>
<td>1072/1642</td>
<td>573/1325</td>
<td>531/919</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>1091/1495</td>
<td>531/919</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Students</td>
<td>ABCDE/All</td>
<td>ABCDE/All</td>
<td>ABCDE/All</td>
</tr>
<tr>
<td></td>
<td>Base</td>
<td>295/762</td>
<td>311/667</td>
<td>398/567</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>267/634</td>
<td>398/567</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Students</td>
<td>ABCDE/All</td>
<td>ABCDE/All</td>
<td>ABCDE/All</td>
</tr>
<tr>
<td></td>
<td>Base</td>
<td>0</td>
<td>499/1248</td>
<td>310/912</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>0</td>
<td>499/1248</td>
<td>310/912</td>
</tr>
</tbody>
</table>

However, only students assessed both baseline and post instruction were used in this study.
Data Source

The research team piloted the 39 multiple-choice items and employed a Rasch analysis, as Bradley and Sampson (2006), to revise and improve the quality of the student assessment prior to collection. The temperature construct can be partitioned into five subtopics: foundations, properties of matter, energy transfer, phase change and thermal energy. Students were administered only subtopics of temperature concepts for which their teachers reported having taught. Administration of the student assessments occurred at the beginning of the academic semester prior to any instruction. The teachers were asked to administer the same assessment within at least two weeks after the temperature unit was covered to serve as a post instruction measure.

Data Collection Procedure

The experimental design for this project influences the timing of the data because the teachers were partitioned into three cohorts. The first cohort entered the project in the summer of 2006 and immediately had the opportunity to participate in the first NU course covering topics of temperature. The second cohort volunteered to serve as controls first and participate in the course the following year in the summer of 2007. The final cohort of teachers entered as controls during the summer of 2007 followed by the first NU course in the summer of 2008 (see Figure 2).
Teachers were given detailed instructions to assign student identification numbers – comprised of the unique teacher identification number, class period number, and unique student number within the class. Teachers kept record of these identification numbers so that all students remain anonymous throughout the research project. Detailed instructions were also given to teachers regarding to administration of the assessment to students (see Appendices).

Administration of the student assessments occurred at the beginning of the academic year prior to any instruction in August or January depending on traditional or block scheduling. Teachers were asked to administer the same assessment using the same student identification numbers within at least two weeks after the temperature unit was covered.

Prior to scanning the data for electronic storage in an Excel file, the data were checked to ensure its quality. Answer sheet were checked for stray marks and to make sure the answers were bubbled in correctly for scanning. Since teachers had the option of giving a subset of the assessment based on what topics were covered during instruction, two independent checks were in place to guarantee the appropriate key sheet was included for grading the assessment. Students
without meaningful identification numbers, students not completing more than 20% of the assessment, and students with answers bubbled in past the number on the assessment were not deemed as quality data and are not kept in the data set for analysis.

Data Analysis

The data were collected from the NU baseline and post instruction multiple-choice assessment designed to measure middle school science student understanding of temperature. The dichotomous Rasch (1960) model was applied to racked and stacked data matrix, which is represented with \( \log_e \left( \frac{P_{ni1}}{P_{ni0}} \right) = B_n - D_i \). Here, \( P_{ni1} \) and \( P_{ni0} \) are the probability that person \( n \) encountering item \( i \) is observed in category 1 or 0, \( B_n \) is the ability measure of person \( n \), and \( D_i \) is the difficulty measure of item \( i \). WINSTEPS software, version 3.65 (Linacre, 2006) was used for the analysis.

Fit statistics were examined for the racked and stacked analyses to determine fit of the data to the model. Wright and Linacre (1994) recommend an acceptable range of 0.7-1.3 for infit and outfit mean square statistics for multiple-choice assessments. Average person measures and spread of items along the continuum were examined within the racked analysis to determine how the assessment targeted the ability of the examinees. Variable maps for the racked data were examined to determine changes in item difficulties from baseline to post assessment. Student measures for the stacked analysis were examined to determine changes in performance from baseline to post assessment.

Results and Discussion

Experimental and control teachers administered a student assessment to all classes being taught temperature concepts. This analysis includes only students that were assessed both
baseline and post instruction. These data include all of the following categories of students:
students having teachers with no NU course (Year 0), students having teachers with one NU
course (Year 1), and students having teachers with two NU courses (Year 2) (see Table 6).

Table 6

*Number of students with teachers by cohort and number of NU courses*

<table>
<thead>
<tr>
<th>Year 0 (No NU course)</th>
<th>Year 1 (1 NU course)</th>
<th>Year 2 (2 NU courses)</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cohort 1</td>
<td>0</td>
<td>1212</td>
<td>779</td>
</tr>
<tr>
<td>Cohort 2</td>
<td>404</td>
<td>479</td>
<td>0</td>
</tr>
<tr>
<td>Cohort 3</td>
<td>792</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>All</td>
<td>1196</td>
<td>1691</td>
<td>779</td>
</tr>
</tbody>
</table>

According to Wright (2003), there are two ways to structure the data matrix and extend
the Rasch model to measure change in understanding over time provided the data fit the model.
By racking the data, the Rasch model can be used to examine changes in item difficulty
estimates over time. On the other hand, stacking the data within the Rasch model allows the
investigation of changes in person estimates over time. These two methods of analysis answer
very different research questions, one focusing on changes in student performance while the
other focuses on changes in content understanding over time. By applying the Rasch model to
the racked and stacked data, interpretations can be made regarding change in student
performance and in item functioning over time, even at the individual level.

Prior to any interpretations, inspecting fit of the data to the model is essential. The
separation and reliability for the persons and items were examined laying the foundation for the
The person and item separation for the racked and stacked data are summarized (see Table 13). The item separation for the racked data was acceptable (greater than 2) for the purpose of this analysis (Wright & Stone, 2004). The person separation for the stacked data was less than 2 for both analyses.

<table>
<thead>
<tr>
<th>Item Separation</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Racked</td>
<td>1.63</td>
</tr>
<tr>
<td>Stack</td>
<td>1.27</td>
</tr>
</tbody>
</table>

Table 13

Person and item summary statistics for racked and stacked data

<table>
<thead>
<tr>
<th>Person Separation</th>
<th>Person Reliability</th>
<th>Item Separation</th>
<th>Item Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Racked</td>
<td>1.63</td>
<td>23.11</td>
<td>1.00</td>
</tr>
<tr>
<td>Stacked</td>
<td>1.27</td>
<td>32.15</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Such a low person separation was not a surprise considering there are only 39 items to separate a very large number of students. In addition, the person separation and reliability was not expected to be high considering all students are in similar regions of rural Appalachia forming a homogeneous group of middle school science students.

The underlying assumption of the Rasch model is unidimensionality, referring to measuring only one attribute at a time which in this study is temperature understanding (Bond & Fox, 2001). The construct can be partitioned into five subtopics encompassed by understanding of temperature: foundations (1-14), properties of matter (15-17), energy transfer (18-26), property changes (27-32), and thermal energy (33-39). A principal component analysis confirmed that the five concepts making up the construct of temperature understanding did not violate the assumption of unidimensionality. Empirical fit of the data to the model was inspected for the stacked and racked data to ensure valid interpretations of the NU data using these analyses. In examining infit and outfit mean square statistics exceeding 1.3 (Wright & Linacre,
no items were highlighted within the racked analysis as misfitting. The only item resulting in a negative point measure correlation for the racked analysis is baseline item 6, which indicates students estimated with higher levels of understanding of temperature are unexpectedly missing this item. Fit of the data to the model must also be examined to make interpretations using person estimates. For the stacked data, 506 students out of 7332 resulted in high fit statistics. The response patterns for these misfitting students will be further investigated according to the misfit diagnosis guidelines of Wright and Linacre (1994a).

Stacked Results and Discussion

A primary interest of researchers is typically to determine whether the NU teacher training course impacted the change in student understanding. To address this question, the data were stacked to explore changes in student performance from baseline to post assessment. Here the Rasch model was employed to calibrate the raw scores onto the same ruler under consideration for fit of data to the model, followed by analysis of variance to compare results across control and experimental students. The interpretations are then focused to results unique to the application of the Rasch model.

The stacked analysis allows the comparison of baseline and post student measures on the same performance continuum. The person estimates were rescaled to resemble the raw score scale of 0-39 questions to ease the presentation of the Rasch results while maintaining the properties of the newly calibrated logit measures. The expectation was for person estimates to increase from baseline to post assessment which indicates students performed better on the assessment after instruction. Descriptively, 2394 out of 3666 students resulted in a higher person estimate from baseline to post assessment. Of these, 1704 students were experimental students; furthermore, 552 students of 779 had teachers that participated in two NU course while 1152
students of 1691 had teachers participating in one NU course. An analysis of variance on these calibrated estimates revealed a significant increase in the average change in student performance estimates for experimental students compared to control students (see Table 14).

Table 14

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std.</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Exp</td>
<td>1691</td>
<td>17.189</td>
<td>2.074</td>
<td>10.39</td>
<td>15.69</td>
<td>17.27</td>
<td>18.69</td>
<td>23.90</td>
</tr>
<tr>
<td>Y0</td>
<td>17.072</td>
<td>2.068</td>
<td>9.27</td>
<td>15.69</td>
<td>17.04</td>
<td>18.29</td>
<td>26.90</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>17.094</td>
<td>2.161</td>
<td>9.93</td>
<td>15.81</td>
<td>17.18</td>
<td>18.46</td>
<td>24.65</td>
<td></td>
</tr>
<tr>
<td>Post Exp</td>
<td>1691</td>
<td>18.635</td>
<td>2.782</td>
<td>9.10</td>
<td>16.77</td>
<td>18.46</td>
<td>20.50</td>
<td>31.80</td>
</tr>
<tr>
<td>Y0</td>
<td>18.583</td>
<td>2.820</td>
<td>10.40</td>
<td>16.78</td>
<td>18.29</td>
<td>20.10</td>
<td>38.10</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>17.785</td>
<td>2.467</td>
<td>7.80</td>
<td>16.24</td>
<td>17.77</td>
<td>19.53</td>
<td>27.89</td>
<td></td>
</tr>
<tr>
<td>Post-Baseline</td>
<td>Exp</td>
<td>1691</td>
<td>1.447</td>
<td>-6.30</td>
<td>-0.20</td>
<td>1.41</td>
<td>3.14</td>
<td>11.80</td>
</tr>
<tr>
<td>Y0</td>
<td>1.511</td>
<td>2.525</td>
<td>-7.20</td>
<td>0.00</td>
<td>1.42</td>
<td>3.12</td>
<td>11.30</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>0.691</td>
<td>2.442</td>
<td>-9.80</td>
<td>-0.70</td>
<td>0.60</td>
<td>2.30</td>
<td>9.90</td>
<td></td>
</tr>
</tbody>
</table>

Assuming control and experimental teachers are similar with the exception of participating in the NU teacher training course, a much larger percentage of students increased their performance after instruction implying the professional development ultimately impacted student learning of temperature. While these results and interpretations can be made using raw scores and an analysis of variance, the Rasch model followed by analysis of variance provides accurate results for these data in that students with the same raw score can be distinguished based on individual item responses and that missing data is not treated as incorrect since not all students were given the entire assessment.
Although the use of the Rasch model coupled with analysis of variance provides appropriate results for these data, this method of analysis extends the conversation beyond traditional interpretations. With individual person and item standard error estimates, interpretations can go beyond the aggregate data and include qualitative, individual student interpretations. Individual student performance estimates with its own standard error enables the discussion of changes in individual student performance in relation to other students, items, or proficiency levels along the continuum. Another powerful attribute of the Rasch model is the estimation of items and persons invariant of the other but calibrated on the same ruler enabling the conversation of person estimates compared to items on the assessment. From the person and item map, interpretations can be made about how the assessment is targeting student understanding of temperature over time. These interpretations can be used by teachers to inform their instruction, at the individual student or classroom level, as well as curriculum developers of the NU teacher training course.

Recall that the items assessing five concepts are encompassed by the construct of understanding of temperature: foundations (F), properties of matter (PM), energy transfer (ET), property changes (PC), and thermal energy (TE). Interpretations regarding the relation between the students and the items on the assessment can be made using the item and person map within the stacked analysis. The item mean is higher than the person mean which indicates the assessment in its entirety was too difficult for the students. Six of the fourteen foundations items are estimated at or below the student mean. Gaps in the items reveal where the assessment is not spanning the student performance continuum adequately. Notice gaps occur in between the item and person means and one standard deviation below the person mean (see Figure 6).
Figure 6. Item Map of Stacked Analysis

```
<more>|<rare>
32          .  +
           .  |
31          .  +
           .  |
30          .  +
           .  |
29          .  +
           .  |
28          .  +
           .  |
27          .  +T
           .  |
26          .  +
           .  |
25          .  +  PC27
           .  |  TE37  F4  F6
24          .  +  ET25
           .  |  S  TE36  F8
23          .#  +  ET24
           .#  T|  F12  TE34
22          .#  +  ET21  ET23  PC29  TE39
           .##  |  PM17
21          .##  +  PM15  ET19  ET20  F9
           .###  |  F11  PC30  PC31
20          ###  S+  PC28
           ######  |M  ET22  F7
19          .#####  +
           .#####  |  F10  TE38
18          .#####  +
           .#####  |M  F13
17          .#####  +  F3  PC32  TE35
           .###  |  ET18
16          ###  +S  F2
           .###  |  PM16
15          ###  S+  ET26
           ###  |
14          .#  +
           .#  |  F1  F14  F5
13          .#  T+
           .#  |T
12          .  +
           .  |
11          .  +  TE33
           .  |
10          .  +
           .  |
  9          .  +
           .  |
   8          .  +
<less>|<frequ>
```

EACH '#' IS 71.
Four foundation items (4, 6, 8, and 12) are estimated to be greater than two standard deviations above the person mean which is troublesome if mastery of these concepts is expected to ultimately grasp a proficient understanding of temperature. A way to explore these gaps along the continuum is to consider the possibility that these gaps illustrate proficiency levels in understanding of temperature. Perhaps each gap illustrates a theoretical concept hierarchy and levels of knowledge required to master the understanding of temperature.

Within the stacked analysis, there is an expectation that once students have been instructed on the material they respond significantly different on the baseline assessment compared to the post assessment. However, 17 of the 39 items did not result in differential responding from baseline to post assessment. Specifically, these items were 2, 6, 8, 13, 14, 15, 16, 17, 20, 23, 25, 29, 31, 32, 33, 37, and 38. There are two possible interpretations for this finding. Either the students for the most part answered these items correctly on baseline and post assessment or vice versa. Conversely, the items resulting in differential responding from baseline to post assessment can be interpreted as the topics that overall students learned by the post assessment.

**Racked Results and Discussion**

Limited attention is usually given to investigating specific topics becoming less difficult by the end of the course. Within the racked analysis, item difficulty estimates are inspected and utilized to make interpretations about changes in content rather than student performance. The conversation under these analyses focuses more on the content of the items rather than the general discussion of whether experimental students perform better than control students. For curriculum developers, this information is meaningful in determining potential content areas where the professional development is impacting teachers more effectively and potentially their
students. Unlike person estimates, the goal is for item difficulty estimates to decrease by the post assessment; in other words, the expectation is that once students are instructed on the material the items become easier.

Similar to the findings from the principal contrast analysis, there was no significant difference in the average change in item difficulty estimates from baseline to post assessment across the five topics related to the understanding of temperature. Therefore, students did not perform better overall on any topic from baseline to post assessment. The item map produced by the racked data shows the items from baseline and post instruction assessments overlaid on the same item map allowing examination of changes in distribution of the items along the student performance continuum (see Figure 7).
**Figure 7. Item Map of Racked Analysis**

```
EACH '|' IS 24.
```
The items within the racked data are labeled B1 for baseline assessment item 1 corresponding to P1 for post assessment item 1. Another way to visualize the change in item difficulty estimates is to plot the baseline post-instruction estimates against each other (see Figure 8).

*Figure 8. Scatterplot of item difficulty estimates from racked analysis*

The line ($y = x$) indicating no change in item difficulty estimates has been drawn to illustrate increases or decreases in item difficulty estimates over time. Items above the line represent items estimated as more difficult by the post-instruction assessment while items below the line represent items estimated as less difficult by the post-instruction assessment. If the items only assessed topics covered in the course, the expectation would be that the majority of items would be less difficult by the post instruction assessment (below the line).
Overall the assessment did not target the student ability as well as was hoped – specifically ten baseline items and nine post items were estimated below the person mean. The expectation would be for more items from the post assessment to be estimated as easier after instruction. The distribution of post instruction assessment items shifted downward indicating the assessment was in general less difficult for examinees by the post assessment. Item difficulty estimates for foundational items 4 and 12 decreased by greater than one standard deviation from baseline to post assessment (see Figure 7). Within the racked analysis, six items out of thirty-nine resulted in higher difficulty estimates from baseline to post assessment – namely 3, 7, 10, 11, 34, and 36 (see Figure 7). Considering the first fourteen items on the assessment are considered to be foundations items, it is troublesome that four of these items are more difficult for students on the post assessment. The concern, here, is that students are developing fundamental misconceptions throughout instruction. While caution must be taken with interpretations based on a few items, it is also important to remember the stability of these item estimates.

Linking content of the student assessment with topics from the teacher assessment and the NU teacher training course can inform curriculum developers of content areas in need of more attention. A differential item functioning (DIF) analysis across teachers with student assessment items linked to the teacher assessment or teacher training may lead to significant findings in terms of specific impacts of the professional development. These interpretations can also be appropriate and useful to the participants at the individual classroom level to inform classroom instruction and for teachers to develop instructional strategies based on results.
Limitations

The restructured assessment was not administered again and analyzed with the Rasch model to determine quality of the instrument. The low variability among the response frame also affects the internal validity of this study due to introducing dependency with the stacked data into the model (Wright, 2003). Even though evidence in this analysis suggests an increase in student understanding as a result of the NU distance learning course, neither change in student performance or item difficulty can be absolutely attributed to NU. The lack of random sampling and low variability among the response frame is a threat to the external validity of this study.

Conclusion

In summary, the results from the stacked data indicated students of teachers participating at some level in an NU teacher training course performed significantly better after instruction than students of teachers having no NU course experience. Not only can this analysis provide meaningful interpretations to evaluate the impact of NU on student learning overall, qualitative attributes for student performance at the classroom levels can be useful to teachers in informing instruction. The results from the racked data indicated the majority of items demonstrated a decrease in item difficulty estimates by the post instruction assessment. Characteristics of the items unexpectedly increasing in difficulty by the post assessment should be investigated by the research team to develop potential reasons for this occurrence. Results will be presented to the NU course developers to determine how these results link to the NU teacher training course as well as the teacher assessment.

Educational Importance

The importance of this study is to extend the conversation beyond the overarching question of whether students whose teachers participated in NU performed significantly better on
the post assessment after instruction compared to the students of teachers without NU experience. Instead this study provides a method of analysis encompassing interpretations of this nature as well as others enabling conversations at the individual student or item level. This study illustrates a way to apply the Rasch model to examine change among examinees and items over time in conjunction with traditional statistical analyses in CTT. The racking and stacking of data illustrates an effective and meaningful way to examine assessment data collected at multiple points of time. Locally, researchers involved in the NU course development can use these findings to further improve the course and the student assessment instrument. The interpretations can also be focused at the classroom level where teachers could use these results to inform their instruction of temperature. In a broader sense, researchers across various fields can use the methods shown in this study to improve assessment instruments, course development, and measures of change over time.

References
Liu & W. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 23-44). Maple Grove, MN: JAM Press.


Appendix A: Cover Letter for Student Assessment Instructions

Dear Newton’s Universe Teachers:

As part of the Newton’s Universe project, we are collecting student achievement data on selected science topics addressed in the Newton’s Universe course, Temperature, Heat, and Energy. To accomplish this, we will be administering pre and posttests three times this year. Enclosed in this packet are the Temperature, Heat, and Energy Assessment (THESA) materials for the pretest. It is important that you follow the enclosed instructions to maintain quality and anonymity of the student results.

In this box, you will find the following.

- Instructions for assigning student IDs (blue sheet)
- Instructions for administering & returning the test materials (yellow sheets)
- 1 envelope for each class to be tested. Each envelope contains:
  - Scantron answer sheets
  - Tests
- Classroom set of #2 pencils
- UPS return label and plastic sleeve for sending back all test materials including the answer sheets and all tests. You may re-use the same box.

Please note that the number of envelopes and tests in each box are based upon your responses on the post-course survey. The envelopes are labeled with your own teacher ID number, and the quantity of tests enclosed. Please contact Sally Shafer immediately if the quantities of tests supplied are insufficient.

Please carefully read over the instructions for assigning 5-digit student IDs (blue sheet) and for administering the test (yellow sheet). Before the test day, follow the instructions to assign student IDs to all students enrolled in classes that you teach heat, temperature, and energy concepts. Please be certain to record each 5-digit student ID in a secure place since these IDs will be needed for each test and survey in the project.

On the day of testing, it is important that students use the unique ID numbers you have assigned them. Follow the yellow sheet instructions for test administration, explaining each step that students should complete. Students should not write on the test itself, but should mark their answers on the separate answer sheets.

Please administer the test to students as soon as possible. Testing should be completed no later than September 29, 2006. The deadline for returning all testing materials to Newton’s Universe is October 4, 2006.

If you have additional questions, please feel free to contact the project manager, Sally Shafer. Her contact information is below.

Sally Shafer
177 Chemistry-Physics Building
Lexington, KY  40506-0055
Phone: 859-257-5131
Email: shafer@pa.uky.edu

We appreciate your participation in this project and your assistance in this important partnership with UK and the National Science Foundation.

Sincerely,

Newton’s Universe Staff
Appendix B

DIRECTIONS FOR ADMINISTERING NEWTON’S UNIVERSE TEMPERATURE, HEAT, AND ENERGY STUDENT ASSESSMENT

The Newton’s Universe project is partnering with teachers to improve student understanding of concepts related to temperature and heat. As a participating teacher, your students will be completing the Newton’s Universe Temperature, Heat, and Energy Student Assessment (THESA) three times during the school year.

Complete the blue sheet instructions for ASSIGNING STUDENT IDs before testing day

WHEN YOU ADMINISTER THE TEST, DO SO AS OUTLINED BELOW:

➢ Please administer the exam on a day that most of your class is present (at least 85%). If there is a school function or illness leading to many absences, please postpone the test until you can have most of the class present.

➢ If a student has an IEP for standardized tests, please follow the procedures outlined in their IEP for state standardized assessments.

BEGINNING THE TESTING:

The tests and bubble answer sheets are in unsealed envelopes, one envelope for each science class in which you will be teaching heat and temperature concepts. Do not open testing materials until the time of the test.

➢ Provide each student a scantron (answer sheet) and a #2 pencil (not pen).*
  *Please note: Even though answer sheets explicitly state pen or pencil can be used, the machine used to read the answer sheets will only recognize marks made by a #2 pencil.

➢ Ask each student to fill in their five-digit STUDENT ID in the space provided at the top right corner of the form. They should write their number in the first five boxes starting on the left, filling in the appropriate matching bubbles. Leave the remaining ID boxes blank. It is very important that student IDs are correct, so please double check that students have accurately completed this task. Students should not complete the name or test name spaces on the answer sheet. The section marked For Teacher Use Only should remain blank.

➢ Tell students to not write on the test. To mark an answer they should carefully darken the bubble for that choice on the answer sheet. Please remind students to double check that their answer sheet number matches the test question. Mark only one answer for each question.

➢ Tell students that the last page of the test is a four-part short answer question. They should write their answers, and label them A, B, C & D, in the block provided on the answer sheet.

➢ Hand out the tests. Instruct students to answer all questions and to do their best work.
Appendix B continued

**UPON COMPLETION OF TESTING:**

When a class completes the test, immediately place all the scantron (answer) sheets and tests back in the envelope, including all unused tests and answer sheets. Immediately seal the envelope and write the class number on the label. *It is important that you do not review the students’ responses.*

Do not keep any copies of the test or test questions; return them all to Newton’s Universe.

**RETURNING THE TEST MATERIALS TO NEWTON’S UNIVERSE:**

After all of your classes have finished testing, put all of the sealed envelopes, pencils, and any other testing materials in the shipping package in which they arrived.

Use the pre-printed UPS shipping label to return the test materials to the Newton’s Universe project office. There is an adhesive backed plastic sleeve enclosed for the shipping label. After sealing your box with sturdy tape, attach the label in its plastic sleeve to the outside of the box. Be sure to remove any address or barcode information that might remain on the box from its initial journey to you.

For your convenience, the boxes can be sent directly from your school’s office by calling UPS to arrange for a pickup, and giving them the information that they request from the return label. *If there are other teachers in your school completing testing, they should return their materials separately, in their own shipping package.*