Evaluating and Restructuring Science Assessments: An Example Measuring Student's Conceptual Understanding of Heat

Kelly D. Bradley, Jessica D. Cunningham and Shannon O. Sampson[1]

University of Kentucky

144 Taylor Education Building

Lexington, KY 40507

kdbrad2@uky.edu, jdvirg2@uky.edu, shannon.sampson@uky.edu

[1] All authors contributed equally to this manuscript. Address all inquiries to Kelly D. Bradley, 131 Taylor Education Building, Lexington, KY 40506.

Abstract

When evaluating the quality of any assessment, reliability and validity are critical to the process. This study applied the Rasch model to evaluate the quality of an assessment constructed to measure student knowledge on conceptual understanding of heat. The measure is designed to document the transfer of teacher knowledge attained during a distance learning training unit to their classroom students. Preliminary findings from the Rasch analysis were provided to a research committee and items were modified or replaced for future implementations. Results of the study provide a methodology for constructing quality science education assessment tools and highlight a successful collaboration between science educators and quantitative methodologists.

Evaluating and Restructuring Science Assessments: An Example Measuring Student's Conceptual Understanding of Heat

In an effort to overcome geographic barriers within Appalachia, funding was allocated to develop a unique program of teacher training which combines distance learning with a hands-on, inquiry approach to the instruction of middle school level physical science[2]. The focus of this study was to evaluate and reconstruct an assessment intended to measure students' conceptual understanding of heat. The assessment was constructed to serve as one indication of the transfer of teacher knowledge gained through their training to their classroom students. Here, a one-parameter Item Response Theory model, commonly referred to as the Rasch model, is applied to evaluate the quality of the middle school science assessment. After preliminary analysis, findings are reported to a team of science educators and updates on made to the assessment using a combination of educational theory and quantitative measurement.

Theoretical Framework

Measurement is central to the construction of a quality student assessment, even in the case of classroom-designed or non-standardized assessments. 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)

Although many testing and measurement textbooks present classical test theory as the only way to determine the quality of an assessment (Embretson & Hershberger, 1999), the Rasch measurement model offers a sound alternative to the classical test theory approach. It is based on

[2] *Newton's Universe* funded by the National Science Foundation Grant No. 0437768. Further information can be found at the project website: http://www.uky.edu/NewtonsUniverse.

two fundamental expectations. First, a more able person should have greater probability of success on assessment items than a less able person. Second, any person should always be more likely to do better on an easier assessment item than on a more difficult one. The Rasch model assumes item difficulty is the characteristic influencing person responses and person ability is the characteristic influencing item difficulty estimates (Linacre, 1999). Thus, careful consideration should be given to the construction of assessments. Items should be written clearly, concisely and such that they are not vulnerable to guessing.

In evaluating the quality of instruments and working to reconstruct those instruments, a discussion of reliability and validity is essential. Smith (2004) explains how reliability and various aspects of validity are examined within Rasch measurement. First, reliability is the degree to which an instrument consistently measures what it is intended to measure, or "the degree to which test scores are free from measurement error" (p. 94). To examine reliability, Rasch measurement places person ability and item difficulty along a linear scale. Rasch measurement produces a standard error (SE) for each person and item, specifying the range within which each person's 'true' ability and each item's 'true' difficulty fall. The individual errors are then used to produce a more accurate average error variance for the sample.

Validity is the degree to which an instrument measures what it is intended to measure, which permits appropriate interpretation of scores (Hopkins, 1998). The aspects of validity of measure interpretation, specifically construct and content validity, can be examined within the Rasch measurement framework (Smith, 2004). The foundation of a quality assessment begins with a clear and explicit construct or, dimension that the assessment is intended to measure, known as construct validity. It is also the responsibility of the test writer to transfer the teacher's intentions into items that report students' performance solely based on their intended ability

(Bond & Fox, 2001). The application of Rasch measurement also allows for a review of the content validity of an assessment. Smith (2004) recommends examining where the items fall along the difficulty continuum of the variable. A representative assessment illustrates items spaced evenly along the continuum in addition to items spanning a wide range of difficulties – at least as wide as the range of student abilities. Rasch fit statistics, which are "derived from a comparison of expected response patterns and the observed patterns" (Smith, 2004, p. 103), can be examined to assess the content validity of the assessment. Fit statistics indicate how well the data fit the expectations of the Rasch model, specifically higher performing students should be more likely to answer items of greater difficulty correctly than lower performing students. In this study, the researchers utilized these characteristics to examine the reliability and validity of the *Newton's Universe*[3] student assessment.

Objective

This study applies the dichotomous Rasch measurement model to evaluate the quality of an assessment constructed to measure student *conceptual understanding of heat* and to discuss the restructuring process based upon those results. The investigation begins by providing a framework for the temperature and heat unit, on which the students are being assessed. It will then look at the quality of the assessment beginning with an evaluation of the fit of the data to the model requirements, asking the following questions of the data collected with the *Newton's Universe* student assessment.

1. Are items on the science assessment 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 "conceptual understanding of heat" variable?

[3] For more information on Newton's Universe, see http://www.as.uky.edu/newtonsuniverse/

4. How are the students utilizing the distracters for each item on the assessment? Specifically, are certain distracters triggering misconceptions associated with the science concepts?

Answers to these questions are used to guide the process of restructuring the assessment.

Method

The research team constructed a student assessment comprised of 41 multiple choice items with four answer choices. The student assessment was piloted with a group of middle school students participating in a science camp during the summer of 2006, providing 18 student exams for the calibration process. Rasch measurement models are useful in working with small sample sizes (Wright & Stone, 1979). Although Wright & Stone recommended the goal at least 30 for stable calibrations, a sample of 18 students was the largest number of students available for the pilot. The items on the assessment were categorized into five content domains: foundations, properties of matter, energy transfer, phase change and thermal energy. For all domains, the underlying construct of *conceptual understanding of heat* remains the same; thus, the theoretical framework of unidimensionality is upheld.

Approximately 30 teachers completed the inquiry-based, science distance learning course during the summer of 2006. Participating teachers are expected to teach subsets of the temperature, heat and energy concepts covered in their training. Students in these classes were administered the updated version of the student assessment at the beginning of the academic year. The students are given the assessment again after receiving the specific instruction related to the teacher training concepts and will again be administered the assessment at the end of the academic year. Here, the focus is on the early waves of assessment and the establishment of a reliable and valid assessment.

Data Collection

The student assessment pilot, a multiple-choice assessment designed to measure middle school students' *conceptual understanding of heat*, was administered to eighteen students during a middle school science camp in July 2006. The students will remain anonymous throughout this analysis. Once appropriate revisions have been made to the student assessment, teachers participating in the summer training will administer the pretest at the beginning of the 2006 academic year. Students will be assigned unique student identification number by their teacher so that assessments can be linked over time.

Data Analysis

Data collected from the pilot of the multiple-choice assessment was analyzed using the dichotomous Rasch model, which is represented by $Log_e\left(\frac{P_{ni1}}{P_{ni0}}\right) = B_n - D_i$. 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 3.55 (Linacre, 2005) will be used for the analysis[4].

Fit statistics are used to indicate the extent that responses to each item is consistent with the responses to other items on the assessment (Smith, 2004). Fit statistics indicate how well the data fit the expectations of the Rasch model. In other words, a more able person should have a higher probability of getting any item correct than a less able person (regardless of item difficulty), and a less difficult item should have a higher probability of being answered correctly than a more difficult item (regardless of the ability of the person responding to the item). An accepted cutoff of ZSTD fit statistics between -2 and 2, which indicates the fit statistics are

[4] Code used to run the analysis is available from the contact author. The free student version of Winsteps, the software utilized in this study, is available at www.winsteps.com,

within two standard deviations from the mean of zero (Wright & Masters, 1982), were used to determine item fit in this analysis. Items that do not fit this requirement are reviewed for problematic wording.

The Rasch model produces a difficulty measure for each item on an assessment and an ability measure for each person taking the assessment that are useful in addressing whether the assessment adequately targets the students' abilities. The spread of the items are examined to determine the accuracy of the "yardstick" constructed to measure student knowledge of the concept. When Rasch analysis places items and persons along a "yardstick," one can see where the persons fall (based on their ability) in comparison to where the items fall (based on their difficulty) (Wright & Stone, 2004, p. 31). A well-designed assessment has a distribution of items that is approximately equivalent to the distribution of persons.

Results and Discussion

Student Assessment Pilot Results

The dichotomous Rasch model was applied to the responses of 18 students to the assessment in its original form of forty-one multiple-choice items. First, the item and person separation and reliability were examined prior to any interpretations of the data. The person separation and reliability values for the pilot data were 2.31 and 0.84 respectively. This person separation roughly indicates the number of groups the students can be separated into according to their abilities. Likewise, the item separation and reliability for the pilot data was 1.56 and 0.71. Considering the small sample size, person and item reliabilities are acceptable for the analysis to continue.

All items fit the expectations of the Rasch model with the exception of item 14; in other words, all items except item 14 (ZSTD outfit statistic = 2.0) had ZSTD infit and/or outfit

statistics between -2 and 2. Item 13 was also flagged for review due to a negative point measure correlation. A positive point measure correlation indicates a positive relationship between student performance and correctly answering the item. For this item, one examinee of higher ability answered this item incorrectly while students of much lower ability answered the item correctly. The committee was advised to revisit the distracter in this case to determine if it was triggering a misconception with the learning concept. In a well-targeted assessment, the average person and item measure are approximately equivalent. The pilot data indicated the average item difficulty was slightly above the average student ability, which is expected for a pretest.

Prior to administering the assessment, test developers were asked to provide theoretical difficulty hierarchy of the items. The empirical hierarchy of items was compared to the theoretical hierarchy of items. Difficulty discrepancies between the two hierarchies were examined to determine why items were not functioning as expected. The only major discrepancy between the empirical and theoretical hierarchy of items was the first item on the assessment. Test developers expected this item to be one of the easiest items on the assessment, but empirically it was determined to be one of the most difficult. The committee reviewed this item and determined it was testing a vocabulary term that would become much easier once the students were taught the material.

The item map [on which students are indicated with x's on the left side and items are indicated by their number and specific domain on the right side] was examined for gaps where a number of students were located along the continuum without items targeted at that ability level (see Figure 1 for circles indicating gaps). Inserting items reflecting corresponding levels of difficulty provides more accurate measures of student abilities at these levels. Notice there is a gap between item 28 and items 9 and 12, with three students falling in this ability range.

Similarly, four students fall in the gap existing between item 11 and items 17, 18, 21, 23, 24, 39, and 8. It was suggested to add items at these difficulty levels to provide more precise measures for students at these ability levels.

The item map (see again Figure 1) was also used to examine whether the difficulty of items were spread across all five domains: foundations (F), properties of matter (PM), energy transfer (ET), phase change (PC) and thermal energy (TE). Notice questions from each domain are spread along the continuum. However, four items (18, 21, 23, 24) from the energy transfer domain were located at the same difficulty level. The suggestion was made for the committee to revisit these items to determine if any redundancy existed in the content of these items. Omitting any of redundant items would reduce the amount of time it would take a student to complete the assessment, while the accuracy of the student ability measures would not be affected.

Student responses were also reviewed to determine any distracters in need of revision. If the average person ability measure was not the highest for the correct answer, the item was highlighted for review of distracters. The items flagged for review due to unexpected functioning of distracters include 4, 13, 14, 30, 32, 38, and 40 (see Table 1). In addition, any items with distracters not being used were highlighted for the committee to determine if distracters not used were appropriate answer choices for the item. Items containing distracters not being used be examinees include 2, 3, 6, 12, 29, 31, 35, 36, 37, and 39 (see again Table 1).

Discussion: Student Assessment Pilot Results

Based on the student assessment pilot results, the committee revisited all items flagged for review in the Rasch analysis. The first item on the pilot student assessment was relocated to the fourth item in an effort to place an easier item first on the student assessment. The item flagged for a high outfit ZSTD statistic of 2.0 was reworded because test developers felt students

were overanalyzing the question (see Figure 2 for edits to item 14). The item with the negative point measure correlation (item 13) was deleted because the committee thought the item in general was confusing. Item 18 was deleted from the student assessment since it tested the same concept as item 19, which was revised to make it easier to replace item 18 (see Figure 3 for edits to item 19). Item 23 was removed from the student assessment because the course does not adequately cover the concept tested. No edits were made to items within the properties of matter and phase change domains. A more difficult foundations item was added to increase the span of foundation items along the ability continuum. To fill one gap in the item spread, item 24 was changed to make the question clearer and in turn, less difficult (see Figure 4 for edits to item 24).

The functioning of the distracters was also useful in improving the quality of the student assessment. The answer choices of temperature points were changed to increase the difficulty of the items 12 and 36. For item 3, the fourth answer option was changed because empirically it was not functioning as expected as a distracter. The option was changed to make it a better distracter, which also theoretically increased the difficulty of this item. Since the first answer option was never selected for item 5, this option was altered for the final version of the student assessment. Item 4 was determined to be confusing for many higher ability students. It presents a thermometer being placed in a glass of water for five minutes and then asks students to predict the temperature reading on the thermometer (i.e. same as water, more than water, less than water, or does not have a temperature). The average person measure for the correct answer choice for item 4 was negative, which means the higher ability students did not choose the correct answer. Test developers chose to alter the amount of time the thermometer was left in the water from 5 to 10 minutes because they suspected brighter students thought 5 minutes was not enough time for

the thermometer to reach equilibrium. For the same reason as item 4, adjustments were made to item 40 attempting to improve the functioning of distracters (see Figure 5 for edits to item 40).

Conclusion

Following the Rasch analysis of the pilot data, results were presented to the committee developing the test. The committee reflected on the results and focused on the items that fit the model poorly, reviewing them for clarity. Furthermore, the committee addressed the distribution of items, determining where gaps existed and what additional items might be added to measure that level of understanding. In conjunction with this, the members reviewed how well the assessment was targeted to the ability of the examinees. Special attention was given to the gaps in the spread of items that were located near the corresponding ability level of many respondents. Following the reconstruction process, the committee was asked to develop a new theoretical hierarchy of item difficulty based on the pilot results and any revisions made. When students are administered the baseline assessment in September 2006, the theoretical and empirical hierarchy of items will be compared again to determine if the items are functioning as expected after the revisions were made.

Thorough inspection of all aspects of a student assessment is crucial in providing the most stable, accurate measure of conceptual understanding. The strength of this study is the partnership of science educators in developing the test with researchers in educational measurement to construct a quality assessment. This study provides a model for assessing knowledge transferred to students through teacher training, specifically in the areas of temperature and heat at the middle school level. Findings will support other researchers in attempts to link student performance outcomes to teacher training, classroom teachers

constructing their own assessments and the continued growth of collaborative efforts between the measurement and science education communities.

Figure 1. Persons map of items.

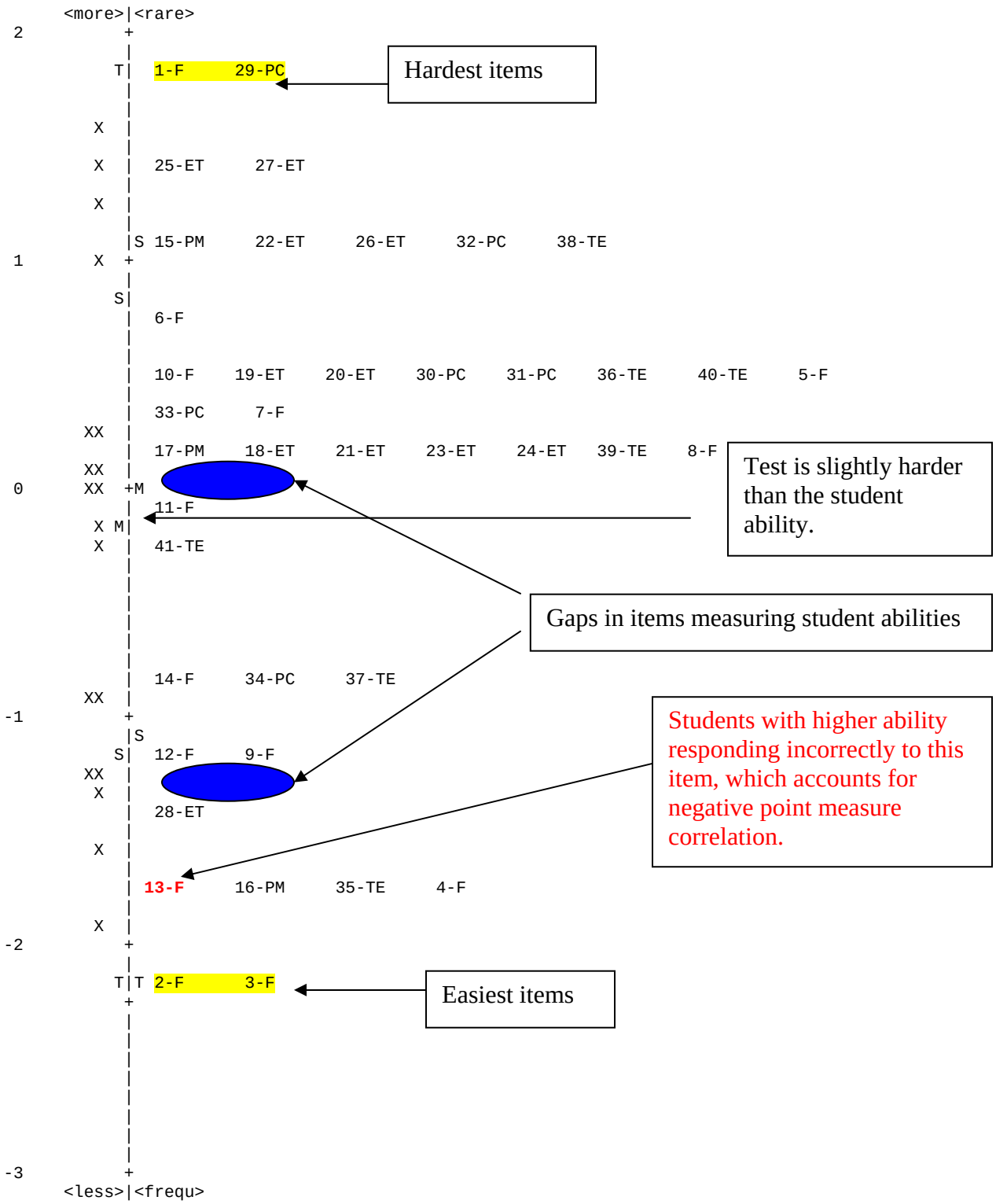


Table 1

Unexpected functioning of distracters for items

Entry Number	Data Code	Score Value	Data Count	Data Percent	Average Measure	Outfit MNSQ	Distracters Not Used
2-F	2	0	1	5	-1.93	0.2	4
	3	0	2	11	-1.07	0.6	
	1	1	16	84	0.03	0.9	
3-F	2	0	3	16	-0.39	2.5	3, 4
	1	1	16	84	-0.16	1.1	
4-F	3	0	1	5	-1.35	0.4	
	4	0	1	5	-0.14	1.4	
	2	0	2	11	0.07	1.7	
	1	1	15	79	-0.16**	1.2	
6-F	2	0	13	68	-0.20	1.4	1, 3
	4	1	6	32	-0.18	1.7	
12-F	4	0	1	5	-1.22	0.4	1
	3	0	5	26	-0.21	1.3	
	2	1	13	68	-0.11	1.4	
13-F	4	0	2	11	-1.14	0.5	
	1	0	1	5	0.11	1.8	
	2	0	1	5	1.55	7.5	
	3	1	15	79	-0.20**	1.2	
14-F	1	0	2	11	-0.59	0.8	
	2	0	3	16	-0.49	2.1	
	4	0	2	11	0.10	3.4	
	3	1	12	63	-0.10**	1.1	
29-PC	1	0	3	16	-1.36	0.2	4
	2	0	13	68	-0.17	1.1	
	3	1	3	16	0.88	0.6	
30-PC	2	0	3	16	-1.40	0.3	
	4	0	6	32	-0.40	0.9	
	3	0	3	16	1.06	3.6	
	1	1	7	37	-0.03**	1.3	
31-PC	3	0	3	16	-1.61	0.2	2
	1	0	9	47	-0.24	1.0	
	4	1	7	37	0.47	0.9	
32-PC	1	0	3	16	-1.04	0.4	
	3	0	7	37	-0.40	0.9	
	4	0	4	21	0.37	1.7	
	2	1	5	26	0.16**	1.6	
35-TE	3	0	3	16	-1.47	0.4	1
	4	0	1	5	-0.02	1.6	
	2	1	15	79	0.05	0.9	
36-TE	1	0	4	21	-0.95	0.5	3
	2	0	8	42	-0.21	1.4	
	4	1	7	37	0.26	1.1	
37-TE	3	0	4	21	-1.02	0.5	1
	4	0	8	42	-0.41	1.4	
	2	1	7	37	0.19	1.1	
38-TE	2	0	4	21	-0.61	0.6	
	1	0	3	16	-0.31	1.7	
	3	0	7	37	0.07	1.5	
	4	1	5	26	-0.16**	1.8	
39-TE	4	0	4	21	-0.96	0.5	1
	3	0	7	37	-0.15	1.3	
	2	1	8	42	0.15	1.5	
40-TE	2	0	3	16	-1.07	0.4	
	3	0	6	32	-0.83	0.6	
	4	0	3	16	0.47	2.8	
	1	1	7	37	0.44**	0.7	

Figure 2. Modifications to item 14.

Pilot Item 14

14. During a 2 week period, Mike read these temperatures on his outdoor digital thermometer before he walked to school. He wore a jacket when the temperature was below 18.3 °C. *How many days did he wear a jacket?*

a. 3
b. 7
c. 8
d. 10

Day	**Temperature (°C)**
Mon. week 1	16.2
Tues. week 1	16.5
Wed. week 1	17.1
Thurs. week 1	16.8
Fri. week 1	17.3
Mon. week 2	17.9
Tues. week 2	18.2
Wed. week 2	18.5
Thurs. week 2	19.2
Fri. week 2	18.1

Revised Item 14

14. During a 2 week period, Mike read these temperatures on his outdoor digital thermometer before he walked to school. If he read a temperature below 18.3 °C, he wore a jacket. On the other days he did not wear a jacket. *How many days did Mike wear a jacket?*

a. 3
b. 7
c. 8
d. 10

Day	**Temperature (°C)**
Mon. week 1	16.2
Tues. week 1	18.2
Wed. week 1	17.1
Thurs. week 1	16.8
Fri. week 1	17.3
Mon. week 2	17.9
Tues. week 2	18.2
Wed. week 2	18.5
Thurs. week 2	19.2
Fri. week 2	18.1

Figure 3. Modifications to item 19

Pilot Item 19

19. *Which of the following is the best example of conduction?*

a. Pavement feels hot against your bare feet.
b. The sunlight feels warm outside during the day.
c. Wind feels cool against your skin.
d. Your hand feels warm above a hot stove.

Revised Item 19

18. *Which of the following is the best example of conduction?*

a. The handle of a spoon in hot soup gets warm.
b. The sunlight feels warm outside during the day.
c. Wind feels cool against your skin.
d. Your hand feels warm above a hot stove.

Figure 4. Modifications to item 24

Pilot Item 24

24. Many people use coolers to keep things cold. *What happens if something hot is placed in the cooler instead?*

a. It stays hot longer.
b. It cools off rapidly.
c. The cooler will not affect its cooling rate.
d. It gets even hotter.

Revised Item 24

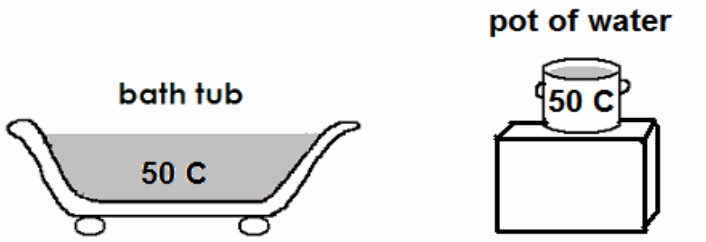
22. On picnics, many people take along a foam cooler to keep things cold. But Cassie wants to keep her fried chicken hot. *What will happen if she puts hot chicken in her cooler instead of anything cold?*

a. The chicken stays hot longer.
b. The chicken cools off quickly
c. The cooler does not affect the chicken's cooling rate.
d. The cooler will cause the chicken to get even hotter.

Figure 5. Modifications to item 40

Pilot Item 40

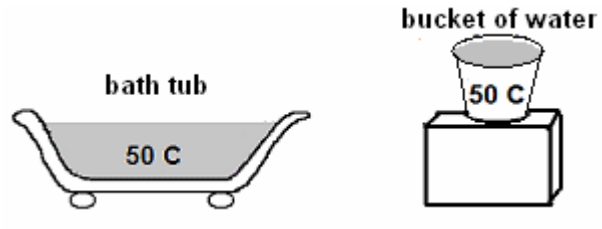
40. When you heat cold water to make it hot, thermal energy must be added to the water. The tub and the pan pictured below contain different amounts of hot water, but they are both at the same temperature, 50°C. *Does it take the same amount of energy to heat the water for each container? Choose the best answer below.*



a. No, the full tub needs more thermal energy.
b. No, the full pot needs more thermal energy.
c. Yes, the tub and the pot both need equal amounts of thermal energy.
d. Not enough information is given.

Revised Item 40

38. When you heat cold water to make it hot, thermal energy must be added to the water. The tub and the bucket pictured below contain different amounts of hot water, but both were filled with hot 50°C water from the faucet. *Does it take the same amount of energy for the water heater to heat the water for each container? Choose the best answer below.*



a. No, the full tub needs more thermal energy.
b. No, the full bucket needs more thermal energy.
c. Yes, the tub and the bucket both need equal amounts of thermal energy.
d. Not enough information is given.

References

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Embretson, S., & Hershberger, S. (1999). *The new rules of measurement*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.) Needham Heights, MA: Allyn & Bacon.

Linacre, J. (1999). *A user's guide to Facets Rasch measurement computer program*. Chicago, IL: MESA Press.

Linacre, J. M. (2005). *WINSTEPS Rasch measurement computer program*. Chicago: Winsteps.com.

Smith, E. (2004). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In E. Smith & R. Smith (Eds.), *Introduction to Rasch measurement* (pp. 93-122). Maple Grove: JAM Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wright, B.D., & Stone, M.H. (2004). *Making measures*. Chicago, IL: The Phaneron Press.