5-20-2013

Measuring Learning Gains in Chemical Education: A Comparison of Two Methods

Thomas C. Pentecost
Grand Valley State University, pentecot@gvsu.edu

Jack Barbera
University of Northern Colorado

JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

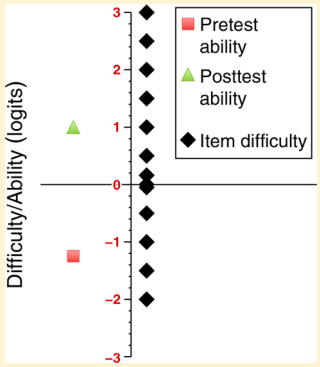
Measuring Learning Gains in Chemical Education: A Comparison of Two Methods

Thomas C. Pentecost*,† and Jack Barbera‡

†Department of Chemistry, Grand Valley State University, Allendale, Michigan 49401, United States

‡Department of Chemistry and Biochemistry, University of Northern Colorado, Greeley, Colorado 80639, United States

ABSTRACT: Evaluating the effect of a pedagogical innovation is often done by looking for a significant difference in a content measure using a pre−post design. While this approach provides valuable information regarding the presence or absence of an effect, it is limited in providing details about the nature of the effect. A measure of the magnitude of the pre−post change, commonly called learning gain, could provide this additional information to chemical education researchers. In this paper, we compare two methods of measuring learning gains using data from large-scale administrations of the Chemical Concepts Inventory at four universities. The intent of this study is to compare various measures of learning gain, not to contrast the teaching effectiveness at the four universities. In this gain analysis, we introduce a method based on Rasch modeling and discuss the advantages offered by this type of analysis over more commonly used measures of learning gain.



KEYWORDS: *Graduate Education/Research, Chemical Education Research, Testing/Assessment*

FEATURE: Chemical Education Research

INTRODUCTION

A laboratory researcher will often design a series of experiments that involve the manipulation of one variable. For example, the rate of a reaction may be measured at one temperature and then again at a different temperature, with all other variables related to the reaction held constant. In this scenario, the researcher is not only interested in whether a change occurs in the rate but also in the magnitude of the change; ultimately, the researcher seeks information to help explain the underlying mechanism from the experimental measurements. A similar situation exists in chemical education research. A researcher might be interested in measuring the change in student learning, analogous to the rate of reaction, based on different instructional methods, analogous to the temperature. While this type of experiment is much more difficult, as holding all other variables constant is difficult, techniques have been developed to aid in the design and interpretation of data from these types of educational experiments.[1,2] Studies of this type assess student knowledge before and after some educational treatment and then evaluate changes in performance for signs of significant differences. This type of comparison informs the researcher about the presence of an effect, but does not provide details about the change. This type of experiment might indicate some sort of dependence (synonymous to a temperature dependence in the analogy), but stops short of providing enough information to relate the effect to the content involved.

What if the educational researcher wants to compare the magnitude of a change in performance from pre- to posttreatment and interpret the change in terms of content? This type of research question is asking about the change in some latent trait due to an intervention. These types of studies seek to qualify learning gains. The measurement of change, learning gains, is not a novel idea within science education[3,4] and has been the focus of many studies in physics and astronomy education research.[5,6] Any database search will quickly reveal numerous studies on student performance prior to and after the implementation of a pedagogical or methodological change within a course. In most studies, the interpretation of the changes has been limited to simply identifying the presence of an effect. It is the goal of this paper to introduce the chemical education community to the common ways of measuring learning gains and introduce a new method that does allow for a more specific interpretation of the changes that have occurred from pre- to post-intervention.

Despite the vast literature on students' pre- and post-implementation performances, the evaluation of learning gains has not been the focus of chemical education research studies. One possible reason for this may be due to the debate in the psychometric community about the ability to accurately measure change. This debate centers on three issues. The first is that *difference scores (post − pre) display a negative correlation with pretest scores.*[3] Therefore, difference scores have a pretest score bias, which means that large positive difference scores are more likely to be observed for students with lower pretest scores. The second issue is that *difference scores often display low reliability.*[7] The reliability of the difference score decreases as the correlation between the pre- and posttest score

Published: May 20, 2013

increases. This presents a problem if the difference score is to be used for a "high-stakes" decision about a student or a pedagogical innovation. In these circumstances the highest reliability possible is required. Therefore, to increase the reliability of the difference requires decreasing the correlation between the pre- and postintervention measures.

This leads to the third problem, the potential for the *"lack of a common trait and scale"*.[8] If the construct being measured changes over time, this will lead to an increase in the reliability of the difference score, due to a decrease in the correlation between the pre- and postintervention measures, but this comes at the cost of validity. This potential problem is often addressed through the use of the same items pre- and postintervention, or with parallel forms of an assessment. However, these precautions do not prevent changes in the construct of the type illustrated by the following example: suppose that initially the items are measuring problem solving, but after instruction they are really measuring recall. The issues related to gain scores are mentioned because it is important for the chemical education community to be aware of the limitations when considering any evaluative technique. A full discussion of these issues is beyond the scope of this communication, whose purpose is to summarize the existing methods of measuring gain that have been used in science education and introduce the chemical education community to a potential alternative method. It is left to individual researchers to form their own opinion of the issue after reading the more detailed discussions[3,8−13] both pro and con.

METHODS OF MEASURING CHANGE

The learning gain or change determined from a pre- versus postexperiment may be calculated several ways. Tornqvist[14] provides an evaluation of several possible calculations. The most obvious manner is a simple difference score, eq 1:

$$\text{Gain} = \text{Post Score} - \text{Pre Score} \tag{1}$$

While this has been used to some extent, issues with this and suggested modifications have been proposed.[7,9] This technique has not found widespread use in science education. An alternative to the simple difference score has been the normalized learning gain, $\langle g \rangle$,[6] eq 2.

$$\langle g \rangle = \left(\frac{\%\langle post \rangle - \%\langle pre \rangle}{100\% - \langle \%pre \rangle} \right) \tag{2}$$

where $\langle\ \rangle$ indicates average scores.

The denominator of eq 2 attempts to accommodate for some of the issues with simple difference scores, namely, ceiling effects and the bias of the absolute gain against high pretest scores. This measure of learning gain has been the most widely used in the physics education community and has had a catalytic role in the revisions of undergraduate physics education at many institutions; a summary of these may be found at the Science Education Initiative Web site.[15] While eq 2 has been widely used, alternatives have been proposed that seek to further correct for pretest score bias.[5,16]

One of the largest, and most influential, studies of learning gain was Hake's investigation[6] of results from the Force Concept Inventory (FCI).[17] Hake reported on the results from 62 introductory physics courses (n = 6542) that administered the FCI both pre- and postinstruction. Courses were deemed "traditional" or "interactive-engagement", based on each instructor's report of their respective teaching methods. From these data, Hake determined average normalized learning gains, $\langle g \rangle$, for each set of courses, finding that the interactive-engagement courses produced higher gain scores. While this study has drawn criticism since its publication,[18,19] it has no doubt shed light on the measurement of learning gains and the impact of various teaching methods. The normalized learning gain has been used as a measure of change in a variety of fields beyond physics, such as astronomy[20] and biology.[5]

Others have recommended an analysis-of-variance-based approach to quantifying learning gains. By using the pretest scores as a covariate for the posttest scores, an ANCOVA analysis does increase the power of the analysis and addresses the potential pretest score bias.[21,22] These approaches have not found widespread acceptance in the science education literature and, like the simple difference and normalized learning gain calculations, the ANCOVA techniques suffer from limitations that arise from being based on an analysis of raw scores.

Because the value of a raw score (e.g., percentage correct) depends upon the difficulty of the test, basing measures of change on raw scores is problematic. The relationship between raw scores and true ability is not necessarily linear and therefore raw scores must be transformed into a true measure before analysis.[22,23] Probabilistic model-based approaches to psychometric measurement have been developed that move beyond the use of raw scores.[24,25] To use these probabilistic approaches, the data must fit a proposed model. Depending on the number of items on the instrument being used, satisfying this requirement can be difficult, as probabilistic models typically require data sets of 200 participants or more. In cases for which the data do not adequately fit the model, the ANCOVA techniques, while not perfect, are an improvement over the simple difference and normalized learning gain techniques.

An alternative to gain calculations based on differences in raw scores would be to use a probabilistic model to estimate student scores. As detailed below, student scores generated with these models are not raw scores based simply on the number (or percentage) of items correctly answered. The Rasch model[26,27] is a probabilistic model that postulates that the observed performance of a student on an item is a function of the student's ability, θ_s, and an item's difficulty, β_i.[28] The functional form of this model is presented in eq 3, where X_{is} is the observed student performance on the item (x = 1 when correct, or x = 0 when incorrect), and the subscripts s and i represent the student and item, respectively.

$$P(X_{is} = x|\theta_s, \beta_i) = \frac{e^{(x(\theta_s - \beta_i))}}{1 + e^{(\theta_s - \beta_i)}} \tag{3}$$

The implication of eq 3 is that if an item's difficulty is greater than a student's ability, the probability of the student responding correctly is low. Another way of expressing this relationship is by taking the logarithm of the ratio of the probability of a correct answer to the probability of an incorrect answer; this is expressed mathematically in eq 4.

$$\ln\left(\frac{P(X_{is} = 1|\theta_s, \beta_i)}{P(X_{is} = 0|\theta_s, \beta_i)} \right) = \theta_s - \beta_i \tag{4}$$

Here the relationship between item difficulty and person ability is more evident. It is important to emphasize at this point an important difference between raw scores and scores generated using probabilistic models, such as Rasch. Raw scores

dx.doi.org/10.1021/ed400018v | J. Chem. Educ. 2013, 90, 839−845

are measures of student ability determined by the fraction of items a student got correct, and are independent of item difficulty. In the probabilistic models, scores are determined by both a student's ability and the item difficulties. Scores generated using a probabilistic model account for the difficulty of items; therefore, these scores are linear. This is not true in the raw score models, for which values at the lower and higher ends of a scale are typically nonlinear. Difference scores (gain scores) are directly interpretable for linear data, as the likelihood of achieving a specific gain score is independent of where on the scale a person begins.

Calibrations of items and persons with the Rasch model put the measures of ability and item difficulty on the same scale, often referred to as the logit (log-odds) scale. Possible values on the logit scale extend from negative infinity to positive infinity. Typically, values fall between −4 and +4, with lower values indicating easier items or lower abilities. For example, the difficulty of item A might be −0.35 logits, and that of item B is 0.75 logits. In this case, item B is harder than item A, because its logit value is larger. A similar comparison can be made between the logit values that represented student abilities. For example, if student A had an ability of −0.35 logits, he or she has a lower ability than student B who has an ability of 0.75 logits. Therefore, student B would have a higher estimate of ability on the construct being measured, and has a higher probability of correctly answering more items than student A. The benefit of putting the item difficulties and person abilities on the same scale does not lie in the logit values themselves, but rather in that the two values, item difficulty and person ability, are directly comparable. Use of the Rasch model in chemistry education research[29−35] is on the rise.

While conventional approaches to determining learning gains have been useful in science education research, their results are often difficult to interpret. Analyzing pre−post changes with the Rasch model offers several distinct advantages. First, this analysis produces truly linear measures of student ability, and these measures may then be subject to conventional statistical analysis.[27] In doing so, the Rasch model fulfills Thorndike's[36] call for the development of techniques to convert observations (scores) into scientific measures. A second advantage of using the Rasch model is the common scale of measure for the item difficulties and student abilities. The advantage this offers to the measurement of learning gains is that a change in ability from pre- to posttest can be directly related to items the student is likely to get correct on the posttest that they were not likely to get correct on the pretest. Therefore, the gain measure is directly interpretable by reference to the item difficulties and their specific content. This analysis can be done at the level of the individual student, a group of students, or at the overall class level.

The match between students' ability scores and assessment items can be seen pictorially through a "Wright map". A Wright map is a vertical plot of the distribution of logit values for both person ability and item difficulty; note that these could be for an individual person or a group of students. Figure 1 shows a hypothesized Wright map for an individual student. The values (3 to −3) are the logit scale; the dashed vertical line separates the person ability data (left side) from the item difficulty data (right side). The ability of the student pre- and posttest is indicated by a square and triangular marker, respectively, and items are typically labeled by their item number; however, in this example they are simply shown as diamond shapes. In Figure 1, the items with difficulties below the student's pretest

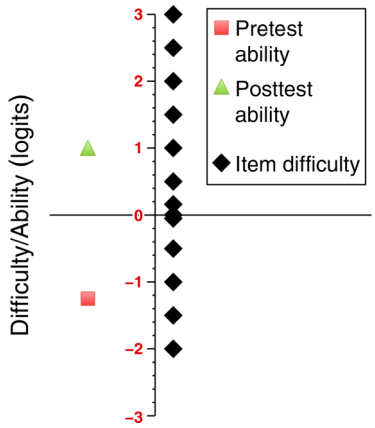


Figure 1. Hypothesized Wright map for an individual student's measures of ability and item difficulty.

ability score are those that the student has a greater than 50% probability of getting correct. As the student's ability increases, a correlation can be made to the items (and hence the content) to which the student is now more capable of correctly responding. Using the Rasch model, gain in student ability between the pre- and postassessments is correlated to students' ability to answer these items. Hence, the gain is interpretable at the item level.

A recent study in astronomy[37] has pointed out the weaknesses of measures based on classical test theory, normalized learning gain calculations, and the advantages of Rasch-based calculations. In the present work, the Rasch learning gain, *RLG*, will be calculated in the same manner (eq 5) used by Wallace and Bailey.[37] This difference calculation is using the Rasch estimate of ability, and not the raw score. This is acceptable here because the Rasch ability estimates represent true linear measures and are suitable for use in a difference calculation[23] where raw scores are not.

$$RLG = \theta_{s,post} - \theta_{s,pre} \quad (5)$$

Prior to using student abilities in eq 5, several preliminary data analyses are required. For clarity, these analyses will be presented using data in the Results and Discussion section.

METHODOLOGY

Instrument

All data were collected using the Chemical Concepts Inventory (CCI).[38] This instrument is designed to assess students' alternate conceptions of chemistry topics typically encountered in high school or first-semester college chemistry courses. The CCI contains 22 multiple-choice items; more than half of the items are in linked pairs, in which the first item probes content knowledge of a specific topic and the second item probes the reasoning for the response. This instrument has been used in studies probing the alternate conceptions of both students[39−41] and instructors.[42] The psychometric properties of the CCI have been evaluated using both classical test theory and Rasch model methodologies.[31] This analysis has shown that the CCI produces valid and reliable data as both a pre- and postinstruction measure of student conceptions.

Population

The CCI was administered during the Fall 2011 semester to students enrolled in a first-semester general chemistry course at four different universities in the United States. Each school gave

dx.doi.org/10.1021/ed400018v | J. Chem. Educ. 2013, 90, 839−845

the exam as both a 30-min pretest and posttest. At each school, the pretest was given within the first two weeks of the course and the posttest was given within three weeks of the final exam. Students in all sections at each school participated and data presented represents only those students who provided consent to do so.

Data Processing

Student responses were gathered using bubble-in response sheets; these sheets were scanned and processed using a spreadsheet program. Data from students who did not provide consent were removed from each data set. Students who did not respond to all 22 items during both administrations were also removed from the data sets. Complete data from 2392 students were obtained. Matched totals for each university are outlined in Table 1. The matched data sets were used to determine the individual ability of each student, as well as the difficulty of the items; average ability levels for each participating university were also calculated.

Table 1. Normalized Learning Gains for the CCI

School (*N*)	⟨Pre%⟩ (SD)	⟨Post%⟩ (SD)	Normalized Learning Gain
A (249)	30.3 (14.0)	35.1 (17.3)	0.06
B (511)	41.2 (17.3)	49.3 (19.8)	0.14
C (416)	40.4 (17.4)	44.6 (18.6)	0.05
D (1216)	44.7 (18.1)	48.0 (18.9)	0.04
All (2392)	41.7 (17.9)	46.4 (19.3)	0.07

Responses from the matched data sets were scored dichotomously. The CCI contains six pairs of two-tier items; these items were only scored as correct if a student responded correctly to both tiers. This scoring methodology reduces the number of items on the CCI from 22 to 16. Collapsing the two-tier items to produce a single score provides a more meaningful interpretation of student understanding, as a student must respond to both tiers correctly. This scoring method is commonly used, including by David Treagust,[43] whose two-tier methodology was used in developing the CCI.

Learning gains were calculated using both the normalized learning gain method and the Rasch learning gain described above. All Rasch analyses were done using Winsteps software.[44] All other change analyses were done using commercially available spreadsheet software. The suitability of the CCI data to a Rasch analysis has been previously established.[31]

RESULTS AND DISCUSSION

Conventional Gain Calculations

Because the normalized learning gain[6] has been the most commonly used measure, our analysis will be limited to this calculation. The gain for each student was calculated and the results averaged. This procedure is essentially equivalent to calculating the normalized learning gain using the average pretest and average posttest scores[45] as in eq 2. Using the criteria established by Hake[6] for the interpretation of ⟨*g*⟩ (high gain, ⟨*g*⟩ ≥ 0.7; medium gain, 0.7 > ⟨*g*⟩ ≥ 0.3; low gain, ⟨*g*⟩ < 0.3), all the learning gains in Table 1 would be classified as low. A typical interpretation of these results is that the normalized learning gain reflects the students' improvement. For example, students at school B had a 14% improvement and overall students improved by 7%. To interpret these gains, one must ask the question: What corresponding content does this 7% represent? On this question the normalized learning gain analysis is silent.

Rasch Learning Gain Calculations

Prior to using this method, the data must be shown suitable for analysis using Rasch techniques. This process, as it relates to CCI data, has recently been published.[31] Once it has been determined that the data are suitable for Rasch analysis, care must be exercised when determining the item difficulties to use between the two time frames. Rasch measures of change must be done with items that are functioning the same in both testing conditions. Failure to check for this will result in change scores that are "spoiled by an uncertain frame of reference".[46] To establish the "constant frame of reference", the analysis should follow the procedure outlined by Wright.[46]

The first step in this procedure is to run a separate Rasch analysis on the student response data pre- and posttest. This yields a set of pretest item difficulties and student abilities, and a set of posttest item difficulties and student abilities. A plot of the posttest item difficulties versus the pretest item difficulties will provide information about the stability of the item parameters over the two administrations. Item invariance is a fundamental assumption of the Rasch model, and items that do not fall on the identity line should be flagged for further analysis. The pre- and posttest item difficulties are plotted in Figure 2 along with an identity line. The identity line is not a line of best fit to the data; rather, it indicates equal *X* and *Y* coordinates across the plot. From the plot in Figure 2, item 5 could be an outlier with respect to the identity line; this indicates that it may not be functioning the same in both frames of reference.

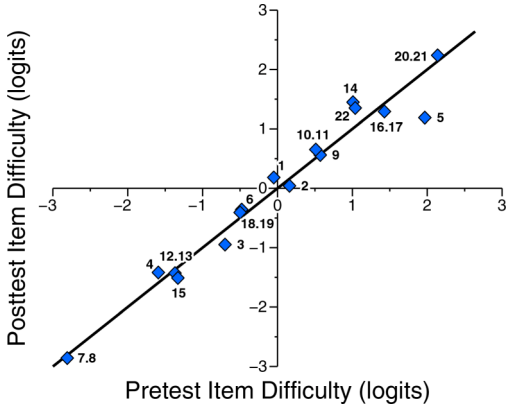


Figure 2. Chemical Concepts Inventory item difficulties from separate pretest and posttest Rasch analysis. Note that the identity line is not a line of best fit to the data; rather, it indicates equal *X* and *Y* coordinates across the plot.

A residual analysis confirms that item 5 is a statistical outlier in the data set. While the other items are not considered outliers, they do not fall exactly on the identity line, therefore, further analysis is required. This next step involves "stacking" the data. Even if item 5 had not looked suspicious, the stacking step would still be a useful step in the analysis because it would provide an additional check on the suitability of the data for the pre−post change analysis.

To stack the data, a new data set is created, with each respondent appearing twice, once pretest and once posttest, and each item appearing once. Therefore, the new data set contains each person's responses pre- and posttest to each item.

dx.doi.org/10.1021/ed400018v | J. Chem. Educ. 2013, 90, 839−845

Items from the first step of the analysis that are truly problematic will show increased misfit to the Rasch model in the stacked analysis. Misfit is a measure of how well the observed data fits the Rasch model.[26] The more "misfit" displayed by an item, the more problematic the item is and a decision should be made about how to handle these items; see Wright[46] for available options. In the stacked analysis of the CCI data, no items displayed misfit to the Rasch model. This indicates that the decrease in difficulty of item 5 from pretest to posttest in Figure 2 likely is due to an instructional effect, not to an issue with the item itself. Therefore, neither item 5, nor any of the other items, requires modification or deletion prior to the gain analysis.

Now that the items have been evaluated for potential problems, the pre- and posttest ability estimates from the Rasch analysis can be used to investigate how the students have changed. This can be useful information; however, as the item difficulties estimated in the stacking analysis represent an intermediate frame of reference, somewhere between pre- and posttest, it is recommended to use this stage only as a check on item functioning, and to move on to an anchoring stage to investigate changes in student ability.

In the anchoring stage, item difficulty data, either pre- or posttest, are used to estimate student abilities. In most instances we are interested in the change or growth of student ability from pre- to postassessment. To do this, we anchor the analysis by using the pretest item difficulties to estimate posttest abilities. In doing this, the gain analysis becomes a direct comparison between a student's pretest and posttest ability and is not conflated with changes in item difficulties. The pretest item difficulties used for our analysis are presented in Table 2.

Table 2. Pretest Item Difficulties Used To Estimate Posttest Student Abilities

Item	Item Difficulty, (β) Logit Units
1	−0.05
2	0.16
3	−0.70
4	−1.59
5	1.97
6	−0.48
7.8	−2.81 (easiest item)
9	0.57
10.11	0.51
12.13	−1.37
14	1.01
15	−1.33
16.17	1.43
18.19	−0.50
20.21	2.14 (hardest item)
22	1.04

Before going on, we briefly interpret the logit values in Table 2. Item combination 7.8 (items 7 and 8 on the CCI are one of the two-tier sets) has the smallest, most negative, logit value, which indicates that it is the easiest item on the instrument, while item combination 20.21 has the largest logit value, indicating that it is the hardest item on the instrument.

The pretest item difficulties (Table 2) can be placed into an input file used to evaluate the students' responses collected during the postsemester administration. This anchoring fixes the item difficulty and evaluates student responses for ability estimates. Without this step, all pre−post analysis results—whether conducted using classical test theory, Rasch, or item response theory methods—are combinations of changing student abilities as well as changing item difficulties.

The results of the Rasch analysis are displayed in two ways. Table 3 shows the numerical results from the analysis and

Table 3. Pretest and Posttest Average Rasch Student Ability Estimates, SD Values, and Rasch Learning Gain Values

School (N)	$\langle\theta_{pre}\rangle$ (SD)	$\langle\theta_{post}\rangle$ (SD)	Rasch Learning Gain $\langle\theta_{post}\rangle - \langle\theta_{pre}\rangle$
A (249)	−1.22 (0.97)	−0.93 (1.1)	0.29
B (511)	−0.53 (1.1)	−0.02 (1.3)	0.51
C (416)	−0.58 (1.1)	−0.31 (1.2)	0.27
D (1216)	−0.30 (1.1)	−0.09 (1.2)	0.22
All (2392)	−0.50 (1.1)	−0.20 (1.2)	0.30

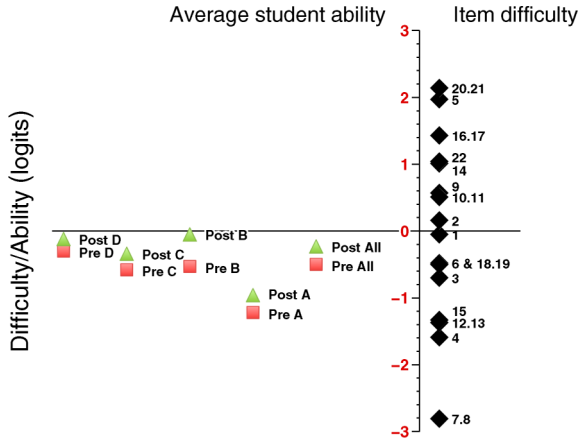


Figure 3. Average abilities from the Chemical Concepts Inventory pretest (red squares) and posttest (green triangles), compared to item difficulties.

Figure 3 presents the results graphically. Student abilities (θ) were determined in separate analyses; Rasch learning gain values were calculated using eq 5. Values represented in Table 3 and Figure 3 are average values for each population. We report averages because we are interested in demonstrating the usefulness of the Rasch method for the evaluation of how each school changes, and are not focusing on student-level changes. As the focus of this manuscript is to describe the method of analysis and not to demonstrate the superiority of one school's instructional practices over another, we will not attempt to interpret the causes for any observed differences between the schools. Such inferences would require detailed information about the teaching environments at each institution. The level of information required to make these inferences was not collected as part of this study.

The value added by using a Rasch-based analysis of change is the ability to correlate gains with specific items and content; this is illustrated in Figure 3. For example, it can be seen that on average, students at school B initially had a 50% chance of correctly answering items 6 (pictorial representation of evaporated water) and 18.19 (conservation of mass for rusting of iron) at the beginning of the semester, because the average ability and the difficulty of these items have the same logit

dx.doi.org/10.1021/ed400018v | J. Chem. Educ. 2013, 90, 839−845

value. At the end of the term they had, on average, a much greater than 50% chance of answering these items correctly, and after the semester students have approximately a 50% chance of answering item 1 (definition of conservation of mass) correctly.

Correlating changes in student ability to items and their content is the type of information that only a Rasch analysis of learning gains can provide. While the CCI data collected for this project does not provide rich data regarding changes in student ability, the potential power of Rasch learning gains is evident. Rasch gain analyses could be extended to smaller groups of students based on some target variable (e.g., gender, learning style, pretest percentile) or even to individual students. These analyses could provide instructors and chemical education researchers with additional data regarding the effect of various pedagogical or curricular changes. Rasch learning gain analyses are being used by Mark Wilson[47−49] and his collaborators in several different psychometric evaluations, including one chemistry project[30] from the BEAR Center at UC−Berkley.

CONCLUSION

Many fields use learning gains as the basis for the evaluation of the effect of an educational innovation. Gain analysis using the normalized learning gain calculations can tell us many things about a course, but cannot answer questions about students' learning of specific content. A Rasch-based analysis of learning gains allows a direct correlation to the improvement in abilities and content through the calibration of the items and abilities on the same scale. This analysis can be done at the institutional, classroom, small group, or individual level. In our analysis of data from the CCI, we found a normalized learning gain on the order of 7% for the entire sample. One institution, B, had a normalized learning gain of 14%. Using a Rasch-based analysis, we have identified a similarly large learning gain for institution B, but can also specify what items (and therefore what content) the gain in ability corresponds to.

In addition to an ability to directly interpret learning gains, the Rasch-based analysis places the measurement squarely in the realm of scientific measurement. Mislevy[50] has elaborated on the advantages of a probabilistic model to the study of gain. These advantages emphasize the fundamental measurement properties inherent in the Rasch model that are missing from analyses based on raw scores. If we wish to evaluate educational innovations in a scientific manner, then we must be willing to expend the effort needed to develop measures of learning that are suitable to Rasch analysis. So while the use of Rasch modeling may be more time-consuming, the rewards are worth the effort. There are more advanced Rasch models that can be used to directly model the change in ability in pre- and posttesting situations.[51−53] Applying these models to the current data will be the subject of future communications.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pentecot@gvsu.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors wish to thank the faculty and students at the participating institutions for their corporation in the data collection. We also thank Steve Pulos and an anonymous reviewer for their assistance and comments that improved this manuscript.

REFERENCES

(1) Field, A.; Hole, G. *How To Design and Report Experiments*; Sage Publications, LTD.: Los Angeles, CA, 2003.
(2) Lewis, S. E.; Lewis, J. E. The Same or Not the Same: Equivalence as an Issue in Educational Research. *J. Chem. Educ.* **2005**, *82* (9), 1408−1412.
(3) Bereiter, C. Some Persisting Dilemmas in the Measurement of Change. In *Problems in Measuring Change*; Harris, C., Ed.; The University of Wisconsin Press: Madison, WI, 1963; pp 3−20.
(4) Lord, F. M. The Measurement of Growth. *Educ. Psychol. Meas.* **1956**, *16* (4), 421−437.
(5) Willoughby, S. D.; Metz, A. Exploring Gender Differences with Different Gain Calculations in Astronomy and Biology. *Am. J. Phys.* **2009**, *77* (7), 651−657.
(6) Hake, R. R. Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses. *Am. J. Phys.* **1998**, *66* (1), 64−74.
(7) Lord, F. M. Elementary Models for Measuring Change. In *Problems in Measuring Change*; Harris, C., Ed.; The University of Wisconsin Press: Madison, WI, 1963; pp 21−38.
(8) Linn, R. L.; Slinde, J. A. The Determination of the Significance of Change between Pre- and Posttesting Periods. *Rev. Educ. Res.* **1977**, *47* (1), 121−150.
(9) Cronbach, L. J.; Furby, L. How Should We Measure "Change"—Or Should We? *Psychol. Bull.* **1970**, *74* (1), 68−80.
(10) Goldstein, H. Measuring Changes in Educational Attainment over Time: Problems and Possibilities. *J. Educ. Meas.* **1983**, *20* (4), 369−377.
(11) Willett, J. B. Questions and Answers in the Measurement of Change. *J. Educ. Meas.* **1988−1989**, *15*, 345−422.
(12) Rodosa, D. R.; Willett, J. B. Demonstrating the Reliability of the Difference Score in the Measurement of Change. *J. Educ. Meas.* **1983**, *20* (4), 335−343.
(13) Hake, R. R. Possible Palliatives for the Paralyzing Pre/Post Paranoia That Plagues Some PEP's. *J. Multidiscip. Eval.* **2006**, *6* (November), 59−71.
(14) Tornqvist, L.; Vartia, P.; Vartia, Y. How Should Relative Change Be Measured? *Am. Statistician* **1985**, *39* (1), 43−46.
(15) University of Colorado Science Education Initiative. http://www.colorado.edu/sei/index.html (accessed May 2013).
(16) Marx, J. D.; Cummings, K. Normalized Change. *Am. J. Phys.* **2007**, *75* (1), 87−91.
(17) Hestenes, D.; Wells, M.; Swackhamer, G. Force Concept Inventory. *Phys. Teach.* **1992**, *30* (3), 141−158.
(18) Hake, R. R. Lessons from the Physics Education Reform Effort. *Conserv. Ecol.* **2002**, *5* (2), 28; http://www.consecol.org/vol5/iss2/art28/ (accessed May 2013).
(19) Hake, R. R. Six Lessons from the Physics Education Reform Effort. http://www.physics.indiana.edu/~hake/SixLessonsD.pdf (accessed May 2013).
(20) Brogt, E.; Sabers, D.; Prather, E. E.; Deming, G. L.; Hufnagel, B.; Slater, T. F. Analysis of the Astronomy Diagnostic Test. *Astron. Educ. Rev.* **2007**, *6* (1), 25−42.
(21) Bunce, D. M.; Hutchinson, K. D. The Use of the GALT (Group Assessment of Logical Thinking) as a Predictor of Academic Success in College Chemistry. *J. Chem. Educ.* **1993**, *70* (3), 183−187.
(22) Barbera, J.; Adams, W. K.; Weiman, C. E.; Perkins, K. K. Modifying and Validating the Colorado Learning Attitudes about Science Survey for Use in Chemistry. *J. Chem. Educ.* **2008**, *85* (10), 1435−1439.
(23) Wright, B. D.; Linacre, J. M. Observations Are Always Ordinal; Measurements, However, Must Be Interval. *Arch. Phys. Med. Rehabil.* **1989**, *70* (12), 857−860.

(24) Embretson, S. E. The New Rules of Measurement. *Psychol. Assess.* **1996**, *8* (4), 341−349.
(25) Wright, B. Fundamental Measurement for Psychology. In *The New Rules of Measurement: What Every Psychologists and Educator Should Know*; Embretson, S. E., Hershberger, S. L., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, 1999; pp 65−104.
(26) Bond, T.; Fox, C. M. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*; Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, 2007.
(27) Wright, B. D.; Stone, M. H. *Best Test Design*; MESA Press: Chicago, IL, 1979.
(28) Masters, G. N. *Educational Measurement*; Australian Council for Educational Research: Melbourne, 1996.
(29) Liu, X.; Boone, W. J. *Applications of Rasch Measurement in Science Education*; JAM Press: Maple Grove, MN, 2006.
(30) Claesgens, J.; Scalise, K.; Wilson, M.; Stacy, A. M. Mapping Student Understanding in Chemistry: The Perspectives of Chemists. *Sci. Educ.* **2009**, *93* (1), 56−85.
(31) Barbera, J. A Psychometric Analysis of the Chemical Concepts Inventory. *J. Chem. Educ.* **2013**, *90* (5), 546−553.
(32) Wei, S.; Liu, X.; Wang, Z.; Wang, X. Using Rasch Measurement To Develop a Computer Modeling-Based Instrument To Assess Students' Conceptual Understanding of Matter. *J. Chem. Educ.* **2012**, *89* (3), 335−345.
(33) Liu, X. Elementary to High School Students' Growth over an Academic Year in Understanding Concepts of Matter. *J. Chem. Educ.* **2007**, *84* (11), 1853−1856.
(34) Herrmann-Abell, C. F.; DeBoer, G. E. Using Distractor-Driven Standards-Based Multiple-Choice Assessments and Rasch Modeling To Investigate Hierarchies of Chemistry Misconceptions and Detect Structural Problems with Individual Items. *Chem. Educ. Res. Pract.* **2011**, *12* (2), 184−192.
(35) Scantlebury, K.; Boone, W. J. Designing Tests and Surveys for Chemical Education Research. In *Nuts and Bolts of Chemical Education Research*; Bunce, D. M., Cole, R. S., Eds.; American Chemical Society: Washington DC, 2008; pp 149−169.
(36) Thorndike, E. L. *An Introduction to the Theory of Mental and Social Measurement*; Teachers College, Columbia University: New York, 1904.
(37) Wallace, C. S.; Bailey, J. Do Concept Inventories Actually Measure Anything? *Astron. Educ. Rev.* **2010**, *9*.
(38) Mulford, D. R.; Robinson, W. R. An Inventory for Alternate Conceptions among First-Semester General Chemistry Students. *J. Chem. Educ.* **2002**, *79* (6), 739−744.
(39) Cacciatore, K. L.; Sevian, H. Incrementally Approaching an Inquiry Lab Curriculum: Can Changing a Single Laboratory Experiment Improve Student Performance in General Chemistry? *J. Chem. Educ.* **2009**, *86* (4), 498−505.
(40) Mayer, K. Addressing Students' Misconceptions about Gases, Mass, and Composition. *J. Chem. Educ.* **2011**, *88* (1), 111−115.
(41) Regan, A.; Childs, P.; Hayes, S. The Use of an Interventions Programme To Improve Undergraduate Students' Chemical Knowledge and Address Their Misconceptions. *Chem. Educ. Res. Pract.* **2011**, *12* (2), 219−227.
(42) Kruse, R. A.; Roehrig, G. H. A Comparison Study: Assessing Teacher's Conceptions with the Chemical Concepts Inventory. *J. Chem. Educ.* **2005**, *82* (8), 1246−1250.
(43) Chandrasegaran, A. L.; Treagust, D. F.; Mocerino, M. The Development of a Two-Tier Multiple-Choice Diagnostic Instrument for Evaluating Secondary School Students' Ability To Describe and Explain Chemical Reactions Using Multiple Levels of Representation. *Chem. Educ. Res. Pract.* **2007**, *8* (3), 293−307.
(44) Linacre, J. M. *Winsteps: Rasch Measurement Computer Program*; Winsteps.com: Beaverton, OR, 2012.
(45) Bao, L. Theoretical Comparisons of Average Normalized Gain Calculations. *Am. J. Phys.* **2006**, *74* (10), 917−922.
(46) Wright, B. D. Time 1 to Time 2 (Pre-Test to Post-Test) Comparison and Equating: Racking and Stacking. *Rasch Meas. Trans.* **1996**, *10* (1), 478.
(47) Wilson, M. Cognitive Diagnosis Using Item Response Models. *Z. Psychol.* **2008**, *216* (2), 74−88.
(48) Wilson, M.; Scalise, K. Assessment To Improve Learning in Higher Education: The BEAR Assessment System. *Higher Educ.* **2006**, *52*, 635−663.
(49) Wilson, M.; Sloane, K. From Principles to Practice: An Embedded Assessment System. *Appl. Meas. Educ.* **2000**, *12* (2), 181−208.
(50) Mislevy, R. On Approaches to Assessing Change; Clarification. http://www.education.umd.edu/EDMS/mislevy/papers/Gain/ (accessed May 2013).
(51) Embretson, S. E. A Multidimensional Latent Trait Model for Measuring Learning and Change. *Psychometrika* **1991**, *56* (3), 495−515.
(52) Adams, R. J.; Wilson, M.; Wang, W.-C. The Multidimensional Random Coefficients Multinomial Logit Model. *Appl. Psychol. Meas.* **1997**, *21* (1), 1−23.
(53) Briggs, D. C.; Wilson, M. An Introduction to Multidimensional Measurement Using Rasch Models. *J. Appl. Meas.* **2003**, *4* (1), 87−100.