Applying Rasch Theory to Evaluate the Construct Validity of Brief Electricity and Magnetism Assessment

Lin Ding

School of Teaching and Learning, The Ohio State University, Columbus, OH 43210

Abstract. The Brief Electricity and Magnetism Assessment (BEMA) is a 30-item multiple-choice test, designed to evaluate student understanding of basic electricity and magnetism (E&M) concepts at the introductory physics level. While previous studies have demonstrated its face and content validity, no efforts were made to evaluate the construct validity of this assessment. In the present study, we use Rasch modeling to explore whether or not the BEMA items can collectively measure the same ability (trait)—student basic understanding and application of E&M concepts. Results from item reliability, person reliability, person-item map, and item fit of Rasch modeling show that in general BEMA items, albeit covering a broad range of topics, form a unidimensional construct.

Keywords: construct validity, Rasch model, assessment, electricity and magnetism, introductory physics, BEMA
PACS: 01.40.Fk, 01.40.gf, 01.40.G-

INTRODUCTION

Assessment is an integral part of effective education. It plays an important role in monitoring student academic progress and evaluating instructional effectiveness. In science education and particularly in physics education, a number of research-based assessment tools have been developed, ranging from knowledge-focused concept inventories, general skill-based tests to meta-cognitive and attitudinal surveys. [1] These tools not only are of great value for pedagogical use in classrooms but also are a rich resource for educational researchers to better explore the nature of student learning in science.

Brief Electricity and Magnetism Assessment (BEMA) is among many research-based concept assessments in physics. [2] It was designed to measure student understanding of basic knowledge in electricity and magnetism (E&M) at the introductory physics level. This 30-item multiple-choice assessment covers a variety of topics, all of which are typically encountered in the college-level E&M course. Albeit similar in nature to other concept inventories, BEMA differs in its breadth of content coverage. A typical concept inventory is often focused on a single topic; for example, the topic of force and motion in the Force Concept Inventory (FCI). [3] However, BEMA is intended to broadly include the entire introductory E&M domain. This broad coverage often brings up such a question that to what extent BEMA can be considered as valid in terms of the construct it purports to measure. In other words, do various topics covered in BEMA form a cohesive construct? Or simply put, does BEMA actually measure anything? [4]

In this paper, we seek to use Rasch modeling to analyze the individual items of BEMA. By evaluating the model fitness as well as other measures from the analysis, including reliability, person-item map and item infit/outfit statistics, we attempt to address the heretofore unanswered question concerning the construct validity of BEMA.

In the following, we first briefly introduce the framework of Rasch theory and its application in the development of educational assessment instruments. We then report analysis of BEMA items and the results of Rasch modeling using a large sample of data. Finally, this paper concludes with discussions on the implication of the current work and future research directions along this line.

RASCH MODELING

The Rasch theory in essence is a Stochastic theory; meaning it conceptualizes the outcome of a student's performance on a specific item as a probabilistic function of the student's ability and the difficulty level of that item. [5] Mathematically this function can be expressed as follows:

$$Pr(X = 1|\theta_n, \delta_i) = \frac{e^{\theta_n - \delta_i}}{1+e^{\theta_n - \delta_i}}$$

where θ_n is the n^{th} student's ability, and δ_i is the difficulty level of the i^{th} item. As seen, the result of the function is solely dependent on the relative values of θ_n and δ_i. When a student's ability level is greater than the difficulty of an item ($\theta_n > \delta_i$), the student will have

a more than 50% chance to answer this question correctly ($Pr > 0.5$). However, if the opposite is true ($\theta_n < \delta_i$), then she/he will have a less than 50% chance to answer it correctly ($Pr < 0.5$). Similarly, if the two variables are equal ($\theta_n = \delta_i$), then the probability for the student to correctly answer this question will be exactly 50%.

Given the fact that the relative relations between person ability and item difficulty determine the probabilistic results, the Rasch theory examines the two estimates simultaneously and generates a common metric scale on which both person abilities and item difficulty levels can be juxtaposed. A plot of such a scale is called a person-item map (see Figure 1 for example).

There are several advantages inherent to this scale. First, the person abilities and item difficulties displayed on this scale are at the interval level. [5] This means that not only the magnitudes of these values are meaningful but the distances between them are also quantitatively interpretable. [6] For instance, consider four students whose ability values on the Rasch scale are 4, 3, 2 and 1 respectively. In this case, the difference between 4 and 2 will be exactly the same as the difference between 3 and 1. However, it is not the case in the Classical Test Theory (CTT) where a student's ability is calculated by simply adding the number of questions she/he has correctly answered. [7] So, in the CTT the difference between the two scores 4 and 2 does not necessarily have the same meaning as the difference between the scores of 3 and 1.

Second, the separation of person and item on the Rasch scale allows the estimates of person abilities and item difficulties to be sample invariant. In other words, by using Rasch modeling one can obtain a set of person ability estimates independently of the items used in an assessment. Similarly, the item difficulty estimates generated from the Rasch modeling also are independent of student samples. [5] This invariance affords useful implication for assessment development and analysis. Since any assessment tool at its early stage must go through pilot-testing at a limited scale, applying the Rasch modeling to evaluate the quality of an assessment can circumvent the trouble of finding a group of students representative of the *entire* test audience.

Third, the juxtaposition of person and item on the Rasch scale offers a direct, easy way to check the relative distributions of students and items, allowing an intuitive grasp of the reliability of measurement. Unlike the CTT where only the reliability of an assessment is estimated, the Rasch theory calculates both assessment reliability and person reliability. [7] Here, assessment reliability reveals the replicability of item ordering if these items were given to another group of students with the same size and abilities. Similarly, person reliability indicates the replicability of person placements on the Rasch scale if these students were given a parallel set of items measuring the same construct. Typically, if the spans of person and item distributions on the Rasch scale share a large overlap, then student abilities and item difficulties often can be reliably estimated through each other.

One important aspect of Rasch theory is worth noting; that is, all items should presumably measure the same ability or trait, which is commonly known as unidimensionality. [5-7] It is this ability (or trait) measured by all functioning items that eventually morphs into the construct of the entire assessment. Explicating this construct therefore is a process of evaluating the construct validity of the assessment. For most concept inventories, the construct of an assessment is evident and easy to demonstrate. However, it is less evident to show if there exists a specific construct for assessments that cover a wide range of topics, such as BEMA.

METHODS

According to the Rasch framework, if BEMA contains a demonstrable construct which all items are intended to measure, then the data will likely fit the Rasch model well. Conversely, if what individual items are measuring is so different that by and large these items do not form a meaningful construct, then the data will likely fail to fit the Rasch model.

Based on this theoretical foundation, we analyzed BEMA data through the Rasch theory to evaluate the model fit. Specifically, we looked into a number of Rasch measures, including assessment and person reliability, item and person estimates, and item fit.

We collected a large sample of 684 data points from more than a dozen calculus-based introductory E&M classes over a three-year period. Student participants were college-level science and engineering majors who took E&M as a mandatory course. We first tabulated raw data into a tab-delimited file for initial screening. It was found that one student left blank on most (> 60%) of the questions and hence was eliminated from our data set. Then the remaining data points were analyzed using Rasch modeling.

DATA AND RESULTS

In this section, we report Rasch results on reliability, item-person estimates and item fit.

A. Reliability

Rasch analysis shows that the person reliability of BEMA is 0.77, and item reliability is 0.99. Note that

both reliability indices can range from 0 to 1, with a higher value representing a higher level of replicability. Our results suggest that the placement of students on the Rasch scale by using the BEMA items can be reliably replicated (r_{person}=0.77). Similarly, the ordering of the BEMA items is also of high reliability and hence can be satisfactorily replicated (r_{item}=0.99).

B. Person-item Map

Figure 1 shows an item-person map of the Rasch analysis. The dashed line in the middle represents a section of the Rasch scale continuum, generated by the modeling and ranging from -4 to +3. On the left side of the scale is the person distribution with each # sign representing 7 students. (Each dot stands for less than 7.) These students are ordered according to their model estimated abilities, varying from the lowest value at the bottom to the highest at the top. On the right side of the dashed line is the item distribution. All items are also ordered in terms of their model estimated difficulties, ranging from the lowest at the bottom to the highest at the top.

As seen, the easiest item is Q1 and the most difficult item is Q28. (Q1 measures student understanding of Coulomb's law and Q28 asks for the direction of an electric field induced by a long solenoid with increasing current.) The majority of the BEMA items are located between -1 and +1 of the scale, covering over half of the person distribution. So, the difficulty levels of these items can be rather accurately estimated by the Rasch model.

On the other hand, nearly a third of the students are located below Q8—the second easiest item of all, indicating that nearly all the BEMA items are at a difficulty level higher than these students' abilities. Therefore, to better estimate these students, we may need more easy items.

Overall, the person-item map suggests that although BEMA may be challenging to some students, the Rasch model can match BEMA items with the student participants fairly well.

C. Item Fit

In addition to the person-item map, we further checked the item fit of the BEMA questions. In the Rasch modeling, the commonly reported fit statistics are the mean squares (the average of squared residuals) of each item. There are two types of mean squares: infit mean squares and outfit mean squares. The difference between the two lies in the weight each statistics gives to person scores. The infit statistics gives more weight to persons whose ability levels are close to the item difficulties, whereas outfit statistics

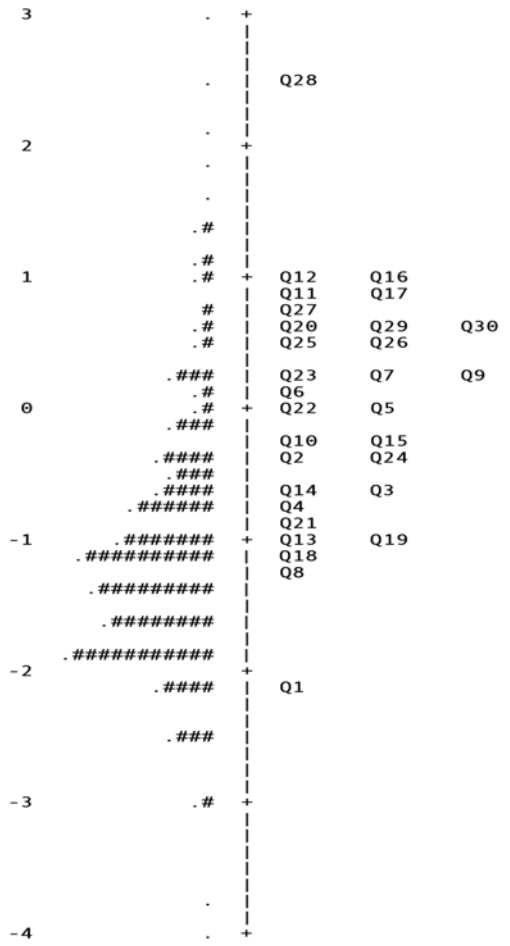

Figure 1 A person-item map for BEMA.

gives equal weight to all persons including outliers. In practical situations, researchers often choose to pay more attention to infit than outfit statistics, since the former is less susceptible to outlying scores.

Both the infit and outfit mean squares can range from 0 to infinity, and ideally the values should be equal or close to 1. For items whose mean square values are greater than 1, they contribute more variations than expected by the model and therefore are called under-fitted items. For those whose mean squares are less than 1, they contribute less variation than expected and hence are called over-fitted items. Both situations are undesirable and can pose a threat to the model fit. Typically, an item is considered to fit well under a unidimensional construct if the infit and/or outfit mean square is in the range of [0.7, 1.3]. [5]

Table 1 displays the fit statistics of BEMA items. As seen, the majority of the items have both satisfactory infit and outfit mean squares, except four items: Q9, Q 11, Q16 and Q 17 (see the shaded cells in Table 1). Q9 requires students to determine a current in salt water in terms of the drift velocity and numbers

of moving charges. Q11 asks students to rank the brightness of identical bulbs in different circuits. Q16 asks students to calculate the potential difference between two points in a uniform electric field. Q17 tests student understanding of electric potential in an open circuit. Among these four items, Q11, Q16 and Q17 show acceptable infit mean squares and thus are less problematic. Only item Q9 has both infit and outfit mean squares beyond the upper limit, indicating that it does not seem to fit under the same construct measured by other items.

In general, the results suggest the existence of a unidimesional construct among the BEMA items despite the fact that these items cover a broad range of E&M topics.

Table 1. Rasch estimated item difficulties and item fit statistics (infit and out mean squares) of BEMA questions.

Item	Diff.	Infit MNSQ	Outfit MNSQ	Item	Diff.	Infit MNSQ	Outfit MNSQ	Item	Diff.	Infit MNSQ	Outfit MNSQ
Q1	-2.1	0.97	0.96	**Q11**	0.84	1.22	1.41	**Q21**	-0.87	0.84	0.8
Q2	-0.44	1.17	1.25	**Q12**	0.94	1.12	1.17	**Q22**	0.05	0.83	0.76
Q3	-0.61	0.95	0.92	**Q13**	-1	0.91	0.87	**Q23**	0.21	0.97	0.94
Q4	-0.81	0.9	0.87	**Q14**	-0.57	1	1.03	**Q24**	-0.34	0.89	0.87
Q5	0.04	0.78	0.71	**Q15**	-0.3	0.77	0.72	**Q25**	0.56	1.13	1.18
Q6	0.12	0.9	0.9	**Q16**	0.99	0.8	0.56	**Q26**	0.46	1.06	1.14
Q7	0.3	1.04	1.04	**Q17**	0.82	1.28	1.39	**Q27**	0.76	1.04	1.22
Q8	-1.28	1.01	1.02	**Q18**	-1.11	1.19	1.26	**Q28**	2.49	0.91	0.81
Q9	0.21	1.43	1.67	**Q19**	-0.95	0.93	0.9	**Q29**	0.61	0.88	0.89
Q10	-0.28	1.06	1.06	**Q20**	0.59	0.86	0.87	**Q30**	0.65	1.08	1.27

CONCLUSIONS AND DISCUSSIONS

While BEMA covers a variety of topics, the individual items in general seem to form a demonstrable construct, which can be captured as the ability of understanding and applying basic introductory E&M concepts. Results from the Rasch analysis of BEMA items support this finding. Specifically, we conducted a multi-year study and collected a sample of 684 data points to examine person reliability, item reliability, person-item map, and item fit of BEMA questions using the Rasch modeling. It is found that the analysis yields satisfactory person and item reliability values. Also, the person-item map displays a fairly good match between students and items, further suggesting a reliable outcome of the Rasch analysis. More importantly, the item fit statistics show that BEMA questions, albeit seemingly measuring different topics, fit well under one unidimensional construct.

This study offers useful implication for assessment development and analysis. A good assessment often has a clear goal by targeting a specific ability (trait) it intends to measure. When an assessment includes a broad range of topics, it usually is challenging to convincingly argue for what exactly it is to measure. So, empirical analysis using Rasch modeling can be an effective way to explicate and confirm the ability (trait) that an assessment purports to measure, and thereby to empirically evaluate the construct validity of the assessment.

ACKNOWLEDGMENTS

The author thanks the three anonymous reviewers for their insightful comments. This study is partially supported by the OSU-EHE SEED grant.

REFERENCES

1. See http://www.ncsu.edu/per/TestInfo.html for a list of assessment instruments.
2. L. Ding, R. Chabay, B. Sherwood, and R. Beichner, Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment, Phys. Rev. ST Phys. Educ. Res. 2 (1), 010105 (2006).
3. D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, The Physics Teacher **30**, 141 (1992).
4. C.S. Wallace and J. M. Bailey, Do concept inventories actually measure anything? Astronomy Education Review 9 (010116-1) (2010)
5. T. G. Bond and C. M. Fox, Applying the Rasch model: fundamental measurement in the human sciences, 2nd ed. (Routledge, Taylor & Francis, N.Y., 2007)
6. W. J. Boone, Townsend, J. S., and Staver, J., Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data, Science Education 95 (2), 258-280 (2011).
7. L. Ding and R. Beichner, "Approaches to data analysis of multiple-choice questions," Physical Review Special Topics - Physics Education Research 5 (020103), 1-17 (2009).