Why Rasch: Selection of a Quantitative Model

Rasch modeling (Rasch, 1980) was the primary quantitative model used to analyze the data collected during the ASPECt project. Here, we describe our rationale for choosing the Rasch model over Classical Test Theory (CTT) and Two- or Three-Item Response Theory (IRT) models.

The Rasch model is a probabilistic model, which facilitates the development, use, and monitoring of robust measurement instruments (multiple choice tests, partial credit tests, Likert scale surveys, and instruments that require judges to assess a person with regard to a trait) (e.g. Boone & Scantlebury, 2006; Liu, 2010; Boone, Staver, & Yale, 2014). Rasch modeling is used worldwide for the construction of measurement instruments in education and medicine (as well as in many other disciplines). Many high-stakes tests have been developed and evaluated using the Rasch model (e.g., international tests such as TIMSS and PISA and statewide tests such as those in California, Ohio, Pennsylvania, Texas, and Illinois). Numerous medical board certification examinations (e.g., National Board of Medical Examiners, American Society of Clinical Pathologist, American Dental Association, American Board of Family Medicine, and Graduate Australian Medical School Admissions Tests) use the model to develop tests, score test results, and link forms of tests to the same invariant ratio scale. Many diagnostic instruments such as the Woodcock-Johnson Test of Cognitive Abilities use the Rasch model for the same reason (Jaffe, 2009).

In the dichotomous Rasch model, the probability that a student will respond to an item correctly is determined by the difference in the student's ability and the difficulty of the item, according to the following equation:

$$\ln\left(\frac{P_{ni}}{1-P_{ni}}\right) = B_n - D_i$$

where P_{ni} is the probability that student *n* of ability B_n will respond correctly to item *i* with a difficulty of D_i (Bond & Fox, 2007; Liu & Boone, 2006). (Note: Rasch modeling uses the term 'ability' to refer to the students' understanding of the science ideas being targeted at the time of testing. It should not be interpreted as an underlying, innate quality of the student, but more narrowly as the students' current understanding of the topic.) An instrument carefully constructed and evaluated with the Rasch model provides a sample-invariant measurement scale.

Limitations of Classical Test Theory. Classical Test Theory (CTT) has several limitations that are addressed by Rasch Modeling. First, in CTT, item difficulties are dependent on the particular students who took the items. Therefore, an item will have a higher difficulty rating when administered to a sample of below-average students than when administered to a sample of above-average students. Likewise, student scores obtained from CTT are test dependent. As a result, changes to a test (e.g. removing or replacing items) may result in an apparent change in a student's performance level. Lastly, often these types of student scores cannot be used in statistical analyses because they may violate the requirements of parametric tests. Rasch modeling overcomes these limitations by providing mutually independent measures of item difficulty and student ability that are expressed on the same interval scale. One major benefit that results is that two different test takers who are evaluated using the Rasch model do not have to complete the identical set of items, and missing data does not require the removal of a test taker from an analysis. This also allows for age-specific forms of an instrument to be designed and performance on each form to be expressed on the same scale.

Rasch Modeling vs. Two- and Three-Parameter Item Response (IRT) Models. The Rasch approach requires that the data fit the model, whereas IRT assumes that the model should fit the data. This means that when using a Rasch approach, items are added and removed from a test instrument to achieve the requirements of the model (unidimensionality, item fit, etc.), unlike IRT modeling, where the researcher chooses a model that best fits the data that result from a given test. In addition, Rasch is a one-parameter model, in which only item difficulty varies. In both the two- and three-parameter IRT, on the other hand, the discrimination index of the items is allowed to vary, and in the three-parameter model, an attempt is also made to estimate guessing. A consequence of varying the discrimination parameter is that item characteristic curves, which are plots that describe the relationship between student ability and the probability of providing a correct response, can cross. When the item characteristic curves for a set of items cross, the order of items from easiest to hardest depends on the student ability (Wilson, 2003). This means that the vertical scale is sample dependent and the students and items cannot be mapped onto a single scale. Because one of the goals of this project was to map the progression of understanding of energy from elementary to high school, we needed a model that would not allow the crossing of these curves.

One criticism of Rasch modeling is the way it accounts for guessing. The fact that guessing is present in multiple choice tests is recognized in the three-parameter IRT model, which adds a pseudo-guessing parameter. Rasch modeling, on the other hand, assumes that guessing adds some amount of random noise to the data. In our item development work, we took both qualitative and quantitative steps to minimize guessing so that the amount of error due to guessing was acceptable. Qualitative steps included (1) incorporating misconceptions into distractors to give students plausible answer choices, (2) ensuring that all of the answer choices relate to the construct being tested, and (3) developing grade-level appropriate versions of the instruments so that students are given items within their ability range.

Rasch Modeling and Learning progressions. The Rasch model is a powerful tool for developing instruments that can be used to validate learning progressions (Wilson, 2009). If the data is shown to have good fit to the model, then the order of item difficulty represents the order in which students develop competency in the ideas being tested (Black, Wilson, & Yao, 2011). Thus, the order of item difficulty can be used to validate the learning progression. Easier ideas are assumed to come earlier in the progression, and more difficult ideas are assumed to come later in the progression. Wright maps (Wilson, 2005) can be generated to visually represent where each idea falls on the scale.

References

Black, P., Wilson, M., & Yao, S.Y. (2011). Road maps for learning: A guide to the navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives, 9*(2–3), 71–123.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (Second ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Boone W. J. & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. *Science Education, 90*, 253-269.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Netherlands: Springer.

Jaffe, L. E. (2009). *Development, interpretation, and application of the W score and the relative proficiency index* (Woodcock-Johnson III Assessment Service Bulletin No. 11). Rolling Meadows, IL: Riverside Publishing.

Liu, X., & Boone, W. J. (Eds.). (2006). *Applications of Rasch Measurement in Science Education*. Maple Grove, MN: JAM Press.

Liu, 2010

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). Chicago: University of Chicago Press (Original work published 1960).

Wilson, M. (2003). On choosing a model for measuring. *Methods of Psychological Research Online*, *8*(3), 1-22.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*(6), 716–730.