Shiken: JALT Testing & Evaluation SIG Newsletter. 13 (3) November 2009 (p. 2 - 5)

Insights in Language Testing: An Interview with Dr. George Engelhard, Jr.

by Phillip Rowles and Joseph Ring

Dr. Engelhard is a Professor of Educational Measurement and Policy at Emory University. He is co-editor of four books, and has authored or co-authored over 100 journal articles, book chapters, and monographs. Four of his most recent co-edited books are *Advances in Rasch Measurement, Volume 1,* and *Objective Measurement: Theory into Practice, Volumes 3, 4, and 5.* He serves on several national technical advisory committees on educational measurement in the USA. This interview was conducted in September 2009.

Could you please tell us about your up-coming lecture series at Temple University Japan?

The purpose of my seminar is to provide an introduction to the concept of invariant measurement. I am currently working on a book on the topic of Rasch models and the quest for invariant measurement. Many of the persistent measurement problems in the human sciences can be addressed in a coherent manner through the lens of Rasch measurement models. My seminar includes an introduction to the Many-Facet Rasch Model (Facets Model), and its use in the development of psychometrically sound assessments. Examples of the topics addressed in this seminar are differential item and person functioning, construction of psychometrically defensible measures, rater biases, equating of assessments, and dimensionality.

How about giving us some insight on your soon to be published book on invariant measurement and the Rasch model?

I am currently working on a book entitled: *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. This book will be published by Routledge (Psychology Press). This book is intended to complement the excellent books on Rasch measurement by Bond and Fox: *Applying the Rasch model: Fundamental measurement in the human sciences*, 2nd Edition and by Wilson: *Constructing measures: An item response modeling approach*. My book grounds Rasch models in the general framework of invariant measurement, and it describes new research on rater-mediated assessments that involve the use of raters with various types of performance assessments.

You are not a stranger to Japan. What interests you about human science measurement in this country?

Japan is part of the great growth and development in the human sciences and measurement theory in the Pacific Rim. There is a need for psychometrically defensible measures of English language proficiency in Japan for admissions to colleges and universities, and Rasch measurement theory provides a useful approach to developing these measures. There are also many research-based applications of Rasch measurement by social scientists in Japan.

What initially interested you in measurement?

When I was an undergraduate student studying psychology and sociology, I came across Paul Lazarsfeld's innovative work on measurement. One of his observations about social science research was that the lack of progress was primarily due to the fact that social scientists did not really know *what* they were measuring. There was no accepted way to tap into the measurement of unobserved latent variables. Good statistical models existed, based on the analysis of variance and regression; however, social scientists were still not clear on how to measure the key latent variables that were being included in their theories. This is not just a statistical problem because measurement is much more than just a simple application of statistics. It also involves interacting with the most basic aspects of social science theories. Therefore, measurement is in fact quite complicated. So, from the time that I was 20 years old I have wanted to focus on measurement, evaluation and statistics, while keeping a close connection to sound social science theories.

There was a strong divide between quantitative and qualitative researchers in the human sciences. What is your view on this divide and what do you think of alternatives proposed by mixed methods researchers?

In my undergraduate years and into graduate school at the University of Chicago, social science researchers were not making a big distinction between quantitative and qualitative research. Rather, the focus was on understanding social and behavioral processes, using multiple methods with triangulation across disciplines. The term "triangulation" to me was a precursor to the popular modern term "mixed methods research." The researcher divide between quantitative and qualitative methods was unfortunate because if we really want to make progress in social and educational research we need access to all possible methods. Mixed methods research has the benefit of trying to build respect for multiple ways of knowing among researchers. We are obtaining data because we all want to solve a similar set of problems, whether it is in education, language testing, or others areas.

This apparent researcher divide is interesting with your insight from the history and socio-political climate of the 1960's and 1970's in the U.S.

Yes, that period in the U.S. was an active time for the evaluation of social programs. This happened as the researcher divide widened. Traditionally, researchers thought you could only work in one methodological area because it is difficult to be competent in one methodology, let alone two. Ultimately, I think we will see more collaboration with teams working together. Personally, I have found it very powerful to use the Rasch model to look at residual analyses, that is, the differences between observed and expected values. These residuals can be examined by content area experts to help us understand why certain people give unexpected responses on certain items. So, I've gone back to the beginning where I'm combining quantitative and qualitative research and measurement. In fact, all measurement starts as qualitative observations. We observe qualitatively and then decide on numerical indices to assign to our qualitative observations. Ultimately, it becomes a cycle going throughout our research and measurement models. We need more qualitative case studies of particular students and areas of research. In other words, we need to collaborate with researchers who have different theoretical and methodological backgrounds to understand more broadly what's going on in the human sciences.

So, either as an individual researcher or collaborating as part of a team, people can reach out and work with others to draw on each other's strengths?

Exactly. I think that this is where future social science research is heading. The publications I have done collaboratively have paid off in many ways. For example, the writing quality of my students at Emory University has improved because of what I've learned about teaching writing. Although, my initial interests were in assessment, my collaboration with writing theorists has lead to a deeper understanding of the interplay between assessment and the teaching of writing. Currently, I am moving into new areas like the assessment of English as a Foreign Language, and I anticipate that I will learn many things about teaching students about the acquisition of a second language while assessing in this new area.

How can we use the Rasch model to build a defensible system in Japanese university entrance examinations?

First, you have to be really clear about the decision and purpose of the assessment. To make things simpler, let's examine one university. When constructing an exam, first define your latent variable, in this case, proficiency in English. Then, you have to think carefully about the exam indicators, observations, and operations that are used to define English proficiency. If the exam is created by a single professor, who changes year-by-year, it creates a problem of maintaining a comparable scoring system over years. How do you know if a single rater's decision isn't based on the luck of the professor draw? For me, fairness is the base of invariance. Invariance means the decision about admitting a student is consistent over years and between professors who are rating the students. So, if a candidate applied for admission in one year and succeeded, she would also be admitted the next year. That might not be the case right now. There are a variety of approaches we can use to make a fairer exam and admission system. For example, it would be useful to have a committee of professors that would be linked with common members from year-to-year. You would have a common group of people who had a shared understanding of what they meant by English proficiency. This committee of 2 or 3 professors would have perhaps a two-year commitment. So, when the new committee came in they could build on the wisdom from previous years. These professors could build up items and item banks for fair year-to-year assessments.

Could you tell us a little more about item banking benefits? Also, could you tell us about anchoring?

Item banking helps when we are trying to develop a variable, like proficiency in English as a Foreign Language. Typically, people tend to focus too much on specific items. They should always try to focus on the latent construct rather than particular items. It should not matter which specific items are used. In the case of university entrance exams, we would maintain the cut-score on that variable using a common subset of items. If there was a 100-item test, as few as 15-items could be maintained as a common subset from the previous year. Of course, security measures would be enforced so that students would not know which items would be repeated. Therefore, item difficulty calibrations could be maintained annually and the selection criterion would be on the latent variable. Equivalency in terms of content between years could be statistically checked using the common item subset to ensure fairness. The creation and maintenance of a fair assessment system is the overall goal when creating an item bank with linked assessors from year to year.

How can we use measurement to make tests fair? Also, what future trends do you see for measurement?

Previously, psychometrics and measurement was perceived as esoteric. When I first started in this field, I initially saw my role as being more theoretical in nature. However, I soon realized that the real challenge was translating and communicating these measurement models in a more practical way. Instead of being a technician, I tried to convey these concepts in a useful, fairer and objective way to wider audiences. One future trend is to recognize that students and items may not function as intended. Studies of differential item and person functioning are increasing our understanding of how to make assessments fair for all students.

What are future trends you see for general education?

Societies sometimes provide unequal opportunities for certain subgroups of people. Educational systems are often the starting point for this unfairness, but they can also be the place for creating fairness. The goal of education is to help all people in the society to have equal access to educational opportunities. Education prepares students for later life. If students have unequal educational experiences, they may struggle to succeed and not be motivated to continue their studies. In the long run, this is harmful not only to particular students but also society as a whole. Societies must fully develop the human capital available to them. If measurement and assessment systems are properly created, then they can play a key role in maintaining fairness. Fairness is something that benefits all of us in a society.

Works Cited

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, G. (in press). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge (Psychology Press).

Engelhard, G., & Wilson, M. (Eds.). (1996). *Objective Measurement: Theory into Practice, Volume 3*. Norwood, NJ: Ablex.

Garner, M., Engelhard, G., Wilson, M., & Fisher, W. (Eds.) (2007). *Advances in Rasch Measurement, Volume One.* Maple Grove, MN: JAM Press.

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

Wilson, M., Engelhard, G., & Draney, K. (Eds.). (1997). *Objective Measurement: Theory into Practice, Volume 4*. Norwood, NJ: Ablex.

Wilson, M., & Engelhard, G. (Eds.). (2000). *Objective Measurement: Theory into Practice, Volume 5*. Stamford, CT: Ablex.