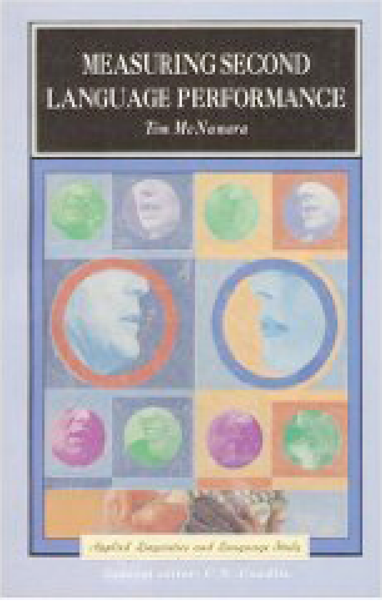
Book Review



Measuring second language performance

by Tim McNamara (1996) ISBN: 0582089077
Harlow, Essex, UK: Addison Wesley Longman Ltd.

Quite soon into the first chapters of this book, I recalled a humorous little adage that often comes to mind in my work; "The truth will make you free, but first it'll make you miserable". The first two chapters present a thorough and critical history of developments in performance assessment as a test method, first in education and occupational settings, then moving on to language testing theory and practice, particularly since the 1970's, when the ideas of communicative competence and authenticity of test-tasks began to be incorporated into discussions about validity in language assessment. McNamara describes language performance assessment as the product of two traditions. The first is the fundamentally pragmatic 'work sample' approach influenced by sociolinguistic theory, which treats the performance itself as the target of assessment (the "strong" argument). The second stems from psycholinguistic theory which views performance as merely a medium and the underlying knowledge and ability as the target (the "weak" argument).

The book's initial message was that language performance tests possess a seductive face validity that obscures the boundaries between what is to be observed, how the subject reacts to the task at hand, and how it is registered as a score. Theories across and within the differing hues in the spectrum from the 'strong' to 'weak' arguments remain far from definitive, and lack empirical evidence to support its constructs. So okay, my most painful doubts during hands-on in-shop experience were not only confirmed, but thanks to the author's comprehensive treatment, were expanded. But in these first chapters, McNamara is simply telling us what we have to face.

McNamara proposes a 'three pronged' approach to tackling the problems he identifies. The first task is to incorporate a model of communicative competence which can explain interaction between all participants in performance assessment, including the interlocutor(s). Secondly, we must direct our research towards finding how significant each variable in our assessment method; (tasks, participants, settings, topics, scales) is to our measurement. Lastly, we must decide, once we have a picture of the impact of these variables, which will fit or inform our model of communicative competence and what the practical boundaries for "testability" are. McNamara uses the remaining two thirds of the book to illustrate how it might be done, and in doing so makes this book a must for those who wish to gain a clear understanding of what Rasch-based analysis can do.

From here on, the book can be read like a detailed journal that an explorer might leave behind for others; telling us what to look for, how to find it, and what it means. McNamara uses as his primary example the development and data analysis of the Occupational English Test (OET), a test of ESL for health professionals in Australia which assessed speaking and writing in work-related simulations. He takes a chapter to explain the procedures for determining the OET's test content (analysis of needs, resources, and the communicative demands of the profession); writing up the specifications, materials and scoring protocols; training and recruiting evaluators; piloting and revision. He includes here, and throughout the book, examples of the actual materials used in the decision making process and in the test itself, all of which can serve as models for the reader to consider.

The final four chapters constitute what I think McNamara meant when he talked of directing our research to find what variables in performance assessment can inform our construct of communicative competence and what cannot. He begins with raters and ratings, presenting evidence of wide variation in how raters apply criteria, even when traditional methods to limit this have been employed. Here the author introduces the advantages of using multi-faceted measurement which, as a Rasch-based method, can process raw scores in such a way to estimate factors such as ability, item (criterion) difficulty, and rater severity all on the same scale. This not only allows the analyst to map the variables side by side to see how they introduce bias or interrelate, but, as demonstrated in this chapter, offers the test user an improved, fairer (more accurate) measurement than raw scores. McNamara then illustrates this with detail and clarity his next chapter on the concepts and procedures of Rasch analysis.

Here begins the most 'cookbook-like' part of the book, resembling Grant Henning's treatment of the subject in his 1987 *Guide To Language Testing*, in that it not only explains how the Rasch model works, but how to interpret the results. What this reporter appreciated most was that here and in subsequent chapters, McNamara uses print-outs and terms which are specific to various Rasch-based computer software that he uses. This allows new users who might otherwise be intimidated by mathematics or psychometric jargon to be on more familiar ground when they try to analyze their own data. The following two chapters present case-studies and reports of related research to build upon the concepts and procedures of Rasch-based analysis introduced earlier.

The author demonstrates how Rasch-based analysis of test data can supply empirical evidence to reveal, at least in part, the impact of various factors, or 'facets' of performance assessment. This fleshes out the 'three-pronged' attack introduced in the beginning of the book with such tools as mapping out a test's purported abilities and skill levels against item difficulty to see how well the rating scale fits the model. Moving more deeply into the area of rating scales and rating criteria, he presents examples of research by himself and others into the impact of individual rater characteristics, descriptors and categories for rating scales, and rater interpretation of the criteria; all of which lead to a deeper, though admittedly unfinished understanding of what performance tests measure.

Perhaps the greatest value this book has for me personally is its last chapter, where I found clear explanations of the more esoteric areas of IRT & latent trait theory and application that had been bothering me for years. These include a breakdown into the various models (Simple Rasch, 2, and 3 parameter models, rating scales-based, partial-credit, and multi-faceted) and their uses. I found NcNamara's treatment of the issue of unidimensionality of data, as well as his explanation of the significance of 'specific objectivity' when choosing the number of parameters for modeling data, both very illuminating and reassuring. The explanations here and throughout the book have also helped me through several user's manuals for IRT-based software such as Quest and Bilog, which offer me a myriad of choices about how to handle data, but don't seem to explain nearly so well how and why I should make them or how I could interpret the results. This book's comprehensive, straightforward and highly readable treatment of L2 performance assessment in general and how useful IRT can be in unlocking its mysteries has made it one of my most treasured sources. I recommend that you get a copy and see for yourself.

- reviewed by Jeff Hubbell

HTML: http://www.jalt.org/test/hub_1.htm / PDF: http://www.jalt.org/test/PDF/Hubbell1.pdf