Language Testing and Evaluation 22

Thomas Eckes

Introduction to Many-Facet Rasch Measurement

Analyzing and Evaluating Rater-Mediated Assessments

PETER LANG
Internationaler Verlag der Wissenschaften

Human ratings are subject to various forms of error and bias. Since the early days of performance assessment, this problem has been sizeable and persistent. For example, expert raters evaluating the quality of an essay, an oral communication, or a work sample, often come up with different ratings for the very same performance. In cases like this, assessment outcomes largely depend upon which raters happen to provide the rating, posing a threat to the validity and fairness of the assessment. This book provides an introduction to a psychometric approach that establishes a coherent framework for drawing reliable, valid, and fair inferences from rater-mediated assessments, thus answering the problem of inevitably fallible human ratings: many-facet Rasch measurement (MFRM). Throughout the book, sample data taken from a writing performance assessment are used to illustrate key concepts, theoretical foundations, and analytic procedures, stimulating the readers to adopt the MFRM approach in their current or future professional context.

Thomas Eckes is Head of the Psychometrics and Research Methodology Department at the TestDaF Institute, University of Bochum. He has taught and published widely in the field of language testing, educational and psychological measurement, and multivariate data analysis. His research interests include rater effects in large-scale assessments, standard setting, and web-based testing.

www.peterlang.de

ISBN 978-3-631-61350-4

Thomas Eckes

Introduction to Many-Facet Rasch Measurement

Analyzing and Evaluating Rater-Mediated Assessments

PETER LANG
Internationaler Verlag der Wissenschaften

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at http://dnb.d-nb.de.

Cover Design:
Olaf Glöckler, Atelier Platen, Friedberg

ISSN 1612-815X
ISBN 978-3-631-61350-4

www.peterlang.de

Introduction to Many-Facet Rasch Measurement

Language Testing and Evaluation

Series editors: Rüdiger Grotjahn
and Günther Sigott

Volume 22

PETER LANG
Frankfurt am Main · Berlin · Bern · Bruxelles · New York · Oxford · Wien

Preface

This book grew out of times of doubt and disillusionment, times when I realized that our raters, all experienced professionals specifically trained in rating the performance of examinees on writing and speaking tasks of a high-stakes language test, were unable to reach agreement in the final scores they awarded to examinees. What first seemed to be a sporadic intrusion of inevitable human error, soon turned out to follow an undeniable, clear-cut pattern: Interrater agreement and reliability statistics revealed that ratings of the very same performance differed from one another to an extent that was totally unacceptable, considering the consequences for examinees' study and life plans.

So, what was I to do about it? Studying the relevant literature in the field of language assessment and beyond, I quickly learned two lessons: First, rater variability of the kind observed in the context of our new language test, the TestDaF (Test of German as a Foreign Language), is a notorious problem that has always plagued human ratings. Second, at least part of the problem has a solution, and this solution builds on a Rasch measurement approach.

Having been trained in psychometrics and multivariate statistics, I was drawn to the many-facet Rasch measurement (MFRM) model advanced by Linacre (1989). It appeared to me that this model could provide the answer to the question of how to deal appropriately with the error-proneness of human ratings. Yet, it was not until October 2002, when I attended a workshop on many-facet Rasch measurement conducted by Dr Linacre in Chicago, that I made up my mind to use this model operationally with the

TestDaF writing and speaking sections. Back home in Germany, it took a while to convince those in charge of our testing program of the unique advantages offered by MFRM. But in the end I received broad support for implementing this innovative approach. It has been in place now for a number of years, and it has been working just fine.

In a sense, then, this book covers much of what I have learned about MFRM from using it on a routine basis. Hence, the book is written from an applied perspective: It introduces basic concepts, analytical procedures, and statistical methods needed in constructing proficiency measures based on human ratings of examinee performance. Each book chapter thus serves to corroborate the famous dictum that "there is nothing more practical than a good theory" (Lewin, 1952, p. 169). Though the focus of the MFRM applications presented herein is on language assessment, the basic principles readily generalize to any instance of rater-mediated performance assessment typically found in the broader fields of education, employment, the health sciences, and many others.

The present book emerged from an invited chapter included in the *Reference Supplement* to the Manual for *Relating Language Examinations to the Common European Framework of Reference for Languages* (CEFR; Council of Europe, 2009), Section H (Eckes, 2009a). Once more, I would like to thank the members of the Council of Europe's Manual Authoring Group, Brian North, Sauli Takala (editor of the Reference Supplement), and Norman D. Verhelst, for helpful comments and suggestions on earlier drafts of that chapter. In addition, I received valuable feedback on the chapter from Rüdiger Grotjahn, Klaus D. Kubinger, J. Michael Linacre, and Carol M. Myford. When the chapter had evolved into this introduction, I was lucky enough to receive again feedback on the completely revised and expanded text, or parts of it, from Mike Linacre and Carol Myford. I highly appreciate their support and encouragement during my preoccupation with some of the more intricate and challenging issues of the MFRM approach. Of course, any remaining errors and shortcomings are mine.

I would also like to express my gratitude to my colleagues at the TestDaF Institute, Bochum, Germany, for many stimulating discussions concerning the design, analysis, and evaluation of writing and speaking performance assessments. Special thanks go to Achim Althaus, Director of the TestDaF Institute, who greatly supported me in striking a new path for designing a high-quality system of performance ratings. The editors of the series *Language Testing and Evaluation*, Rüdiger Grotjahn and Günther Sigott, warmly welcomed my book proposal. Sarah Kunert and Miriam Matenia, research assistants at the TestDaF Institute, helped with preparing the author and subject indexes.

Last, but not least, I would like to thank those persons close to me. My wife Andrea encouraged me to get the project started and provided the support to keep going. My children Laura and Miriam shared with me their experiences of rater variability at school (though they would not call it that), grumbling about Math teachers being unreasonably severe and others overly lenient, or about English teachers eagerly counting mistakes and others focusing on the skillful use of idiomatic expressions, to mention just a few examples. Looking back at my own schooldays, it is tempting to conclude that rater variability at school is one of the most reliable things in life. At the same time, this recurring variability pushed my motivation for finishing the book project to ever higher levels.

Indeed, my prime goal of writing this book was to introduce those who in some way or another employ, oversee, or evaluate rater-mediated performance assessments to the functionality and practical utility of many-facet Rasch measurement. To the extent that readers feel stimulated to adopt the MFRM approach in their own professional context, this goal has been achieved. So, finally, these are times of hope and confidence.

Thomas Eckes
March, 2011

Contents

1

Introduction

This chapter introduces the basic idea of many-facet Rasch measurement. Three examples of assessment procedures taken from the field of language testing illustrate its context of application. The first example refers to a typical reading comprehension test, the second example to a task-based writing performance assessment where raters evaluate the quality of essays, and the third example to rating examinee performance on a speaking test with live interviewers. Having discussed concepts such as *facets* and *rater-mediated assessment*, the methodological steps involved in adopting a many-facet Rasch measurement approach are pointed out. The chapter concludes with a section on the book's purpose and a brief overview of the chapters to come.

1.1 Facets of Measurement

The field of language testing traditionally draws on a large and diverse set of procedures that aim at measuring a person's language proficiency or some aspect of that proficiency (see, e.g., Alderson & Banerjee, 2001, 2002; Bachman & Palmer, 1996; Spolsky, 1995). For example, in a reading comprehension test examinees may be asked to read a short text and to respond to a number of questions or items that relate to the text by selecting the correct answer from several options given. Examinee responses to items may be scored either correct or incorrect according to a well-defined key. Presupposing that the test measures what it is intended to measure (i.e., reading comprehension proficiency), an examinee's probability of getting a particular item correct will depend on his or her reading proficiency and the difficulty of the item.

In another testing procedure, examinees may be presented with several writing tasks or prompts and asked to write short essays summarizing information or discussing issues stated in the prompts based on their own perspective. Each essay may be scored by trained raters using a single holistic rating scale. Here, an examinee's chances of getting a high score on a particular task will depend not only on his or her writing proficiency and the difficulty of the task, but also on characteristics of the raters who award scores to examinees, such as raters' overall severity or their tendency to avoid extreme categories of the rating scale. Moreover, the nature of the rating scale itself is an issue. For example, the scale categories, or the performance levels they represent, may be defined in a way that makes it hard for an examinee to get a high score.

As a third example, consider a face-to-face interview where a live interviewer elicits language from an examinee employing a number of speaking tasks varying in difficulty. Each spoken response may be recorded on tape and scored by raters according to a set of analytic criteria (e.g., comprehensibility, content, vocabulary, etc.). In this case, the list of variables that presumably affect the scores finally awarded to examinees is yet longer than in the writing test example. Relevant variables include examinee speaking proficiency, the difficulty of the speaking tasks, the difficulty or challenge that the interviewer presents for the examinee, the severity or leniency of the raters, the difficulty of the rating criteria, and the difficulty of the rating scale categories.

The first example, the reading comprehension test, describes a frequently encountered measurement situation involving two components or facets: examinees and test items. Technically speaking, each individual examinee is an element of the *examinee facet*, and each individual test item is an element of the *item facet*. Defined in terms of the measurement variables that are assumed to be relevant in this context, the proficiency (or ability, competence) of an examinee interacts with the difficulty of an item to produce an observed response.

The second example, the essay writing, is typical of a situation called *rater-mediated assessment* (Engelhard, 2002; McNamara, 2000), also known as a *performance test* (McNamara, 1996; Wigglesworth, 2008). In rater-mediated assessment, one more facet is added to the set of factors that may have an impact on examinee scores (besides the examinee and task facets)—the *rater facet*. As discussed in detail later, the rater facet is unduly influential in many circumstances. Specifically, raters often constitute an important source of variation in observed scores that is unwanted because it

threatens the validity of the inferences that can be drawn from the assessment outcomes.

The last example, the face-to-face interview, is similarly an instance of rater-mediated assessment, but represents a situation of significantly heightened complexity. At least five facets, and possibly various interactions among them, can be assumed to have an impact on the measurement results. These facets, in particular examinees, tasks, interviewers, scoring criteria, and raters, co-determine the scores finally awarded to examinees' spoken performance.

As the examples demonstrate, assessment situations are characterized by distinct sets of factors directly or indirectly involved in bringing about measurement outcomes. More generally speaking, a *facet* can be defined as any factor, variable, or component of the measurement situation that is assumed to affect test scores in a systematic way (Bachman, 2004; Linacre, 2002a; Wolfe & Dobria, 2008). This definition includes facets that are of substantive interest (e.g., examinees), as well as facets that are assumed to contribute systematic measurement error (e.g., raters, tasks, interviewers, time of testing). Moreover, facets can interact with each other in various ways. For instance, elements of one facet (e.g., individual raters) may differentially influence test scores when paired with subsets of elements of another facet (e.g., female or male examinees). Besides two-way interactions, higher-order interactions among particular elements, or subsets of elements, of three or more facets may also come into play and affect test scores in subtle, yet systematic ways.

The error-prone nature of most measurement facets, in particular raters, raises serious concerns regarding the psychometric quality of the scores awarded to examinees. These concerns need to be addressed carefully, particularly in high-stakes tests, the results of which heavily influence examinees' career or study plans. Many factors other than those associated with the construct being measured can have a non-negligible impact on the outcomes of assessment procedures. Therefore, the construction of reliable, valid, and fair measures of language proficiency depends crucially on the implementation of well-designed methods to deal with multiple sources of variability that characterize many-facet assessment situations.

Viewed from a measurement perspective, an appropriate approach to the analysis of many-facet data would involve the three steps shown in Figure 1.1. These steps form the methodological basis of a measurement approach to the analysis and evaluation of performance assessments, in particular rater-mediated assessments.

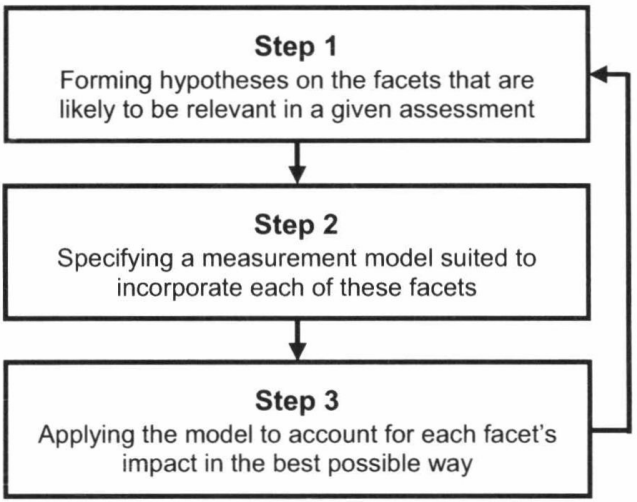

FIG. 1.1 Basic three-step measurement approach to the analysis and evaluation of performance assessments

Step 1 starts with a careful inspection of the design and development of the assessment procedure. Relevant issues to be considered at this stage include defining the group of examinees at which the assessment is targeted, selecting the raters to provide the ratings, and determining the required components of the scoring scheme, such as criteria or scale categories. This step is completed when the factors have been identified that can be assumed to have an impact on the assessment. Usually there is a small set of key factors that are considered on a routine basis (e.g., examinees, raters, tasks). Yet, as explained later, this set of factors may not be exhaustive in the sense that other, less obvious factors could have an additional effect.

Steps 2 and 3, respectively, address the choice and implementation of a reasonable psychometric model. Specifying such a model will give an operational answer to the question of what factors are likely to come into play in the assessment process; applying the model will provide insight into the adequacy of the overall modeling approach, the quality of the measures constructed, and the validity of the conclusions drawn from them. As indicated by the arrow leading back from Step 3 to Step 1, the measurement outcomes may also serve to modify the hypotheses on which the model specified in Step 2 was based or to form new hypotheses that better represent the set of factors having an impact on the assessment. This book deals mainly with Steps 2 and 3.