Language Testing
and Evaluation

22

Thomas Eckes

# Introduction to Many-Facet Rasch Measurement

## Analyzing and Evaluating Rater-Mediated Assessments

PETER LANG
EDITION

# 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 the broader context of its application. In the first example, examinees respond to items of a reading comprehension test. The second example refers to a writing performance assessment, where raters evaluate the quality of essays. In the third example, raters evaluate the performance of examinees on a speaking assessment involving live interviewers. Having discussed key concepts such as *facets* and *rater-mediated assessment*, the general steps involved in adopting a many-facet Rasch measurement approach are pointed out. The chapter concludes with an outline of the book's purpose and a brief overview of the chapters to come.

## 1.1 Facets of measurement

The field of language testing and assessment traditionally draws on a large and diverse set of procedures that aim at measuring a person's language ability or some aspect of that ability (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 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. Assuming that the test measures what it is intended to measure, that is, when the number-correct score can be interpreted in terms of an examinee's reading ability, the probability of getting a particular item correct will depend on that ability and the difficulty of the item.

In another procedure, examinees are presented with several writing tasks and asked to write short essays summarizing information or discussing issues stated in the tasks. 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 ability and the difficulty of the task, but also on various characteristics of the raters, such as individual raters' tendency to assign overly harsh or lenient ratings, or their general preference for using the middle categories of the rating scale. Moreover, the nature of the rating scale itself is an issue. Thus, 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 foreign language face-to-face interview where a live interviewer elicits responses from an examinee employing a number of speaking tasks that gradually increase in difficulty level. Each spoken response is recorded on disk and scored by raters according to a set of distinct criteria (e.g., comprehensibility, content, vocabulary, etc.). In this case, the list of variables that may affect the scores finally awarded to examinees is yet longer than in the writing assessment example. Relevant variables include examinee speaking ability, the difficulty of the speaking tasks, the difficulty (or challenge) that the interviewer's style of interaction 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 ability (or proficiency, competence) of an examinee interacts with the difficulty of an item to produce an observed response (the terms *ability*, *proficiency*, or *competence* will be used interchangeably in this book).

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) or *performance assessment* (Johnson, Penny, & Gordon, 2009; Lane & Stone, 2006). In rater-mediated assessment, one more facet is added to the set of facets 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 (also called graders, markers, scorers, readers, or judges) often constitute an important source of variation in observed (or raw) scores that is unwanted because it threatens the validity of the inferences drawn from assessment outcomes.

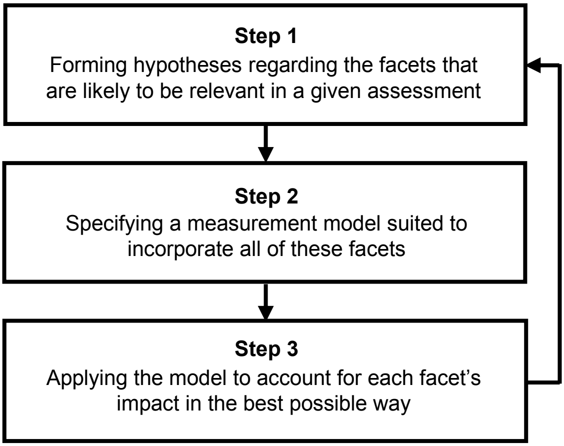
The last example, the face-to-face interview, is similarly an instance of rater-mediated assessment, but represents a situation of significantly heightened complexity. At least five facets, and possibly various interactions among them, can be assumed to have an impact on the measurement results. These facets, in particular examinees, tasks, interviewers, scoring criteria, and raters, in some way or other codetermine 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 or assessment 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, criteria, 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 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 scores in subtle, yet systematic ways.

The error-prone nature of most measurement facets, in particular the fallibility of human 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 assessments, the results of which heavily influence examinees' career or study plans. As discussed throughout this book, many facets 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 examinee ability 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 adequate approach to the analysis of many-facet data would involve three general steps as 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**
Forming hypotheses regarding the facets that are likely to be relevant in a given assessment
**Step 2**
Specifying a measurement model suited to incorporate all of these facets
**Step 3**
Applying the model to account for each facet's impact in the best possible way

Step 1 starts with a careful inspection of the overall design and the development of the assessment procedure. Issues to be considered at this stage include defining the group of examinees at which the assessment is targeted, selecting the raters, and determining the scoring approach (number and kind of scoring criteria, number of performance tasks, scale categories, etc.). This step is completed when the facets have been identified that can be assumed to have an impact on the assessment. Usually there is a small set of key facets that are considered on a routine basis (e.g., examinees, raters, criteria, tasks). Yet, as explained later, this set of facets may not be exhaustive in the sense that other, less obvious facets 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 facets are likely to come into play in the assessment process; applying the model will provide insight into the adequacy of the overall measurement approach, the accuracy of the measures constructed, and the validity of the inferences made from those measures. 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 facets having an impact on the assessment. This book deals mainly with Steps 2 and 3.

## 1.2 Purpose and plan of the book

In this book, I present an approach to the measurement of examinee proficiency that is particularly well-suited to dealing with many-facet data typically generated in rater-mediated assessments. In particular, I give an introductory overview of a general psychometric modeling approach called *many-facet Rasch measurement* (MFRM). This term goes back to Linacre (1989). Other commonly used terms are, for example, *multi-facet(ed)* or *many-faceted Rasch measurement* (Engelhard, 1992, 1994; McNamara, 1996), *many-faceted conjoint measurement* (Linacre, Engelhard, Tatum, & Myford, 1994), or *multifacet Rasch modeling* (Lunz & Linacre, 1998).

My focus in the book is on the rater facet and its various ramifications. Raters are almost indispensable in assessing performance on tasks that require examinees to create a response. Such tasks range from limited production tasks like short-answer questions to extended production tasks that prompt examinees to write an essay, deliver a speech, or provide work samples (Carr, 2011; Johnson et al., 2009). The generic term for these kinds of tasks is *constructed-response tasks*, as opposed to *selected-response tasks*, where examinees are to choose the correct answer from a number of alternatives given. Typical selected-response task formats include multiple-choice or true–false items.

This book heavily draws on a field of application where raters have always figured prominently: the assessment of language performance, particularly with respect to the productive skills of writing and speaking. Since the "communicative turn" in language testing, starting around the late 1970s (e.g., Bachman, 2000; McNamara, 1996, 2014; Morrow, 1979), raters have played an increasingly important role. Yet, from the very beginning, rating quality studies have pointed to a wide range of rater errors and biases (e.g., Guilford, 1936; Hoyt, 2000; Kingsbury, 1922; Saal, Downey, & Lahey, 1980; Wind & Engelhard, 2013). For example, it may be known that some raters tend to assign lower ratings than others to the very same performance; when these raters are to evaluate examinee performance in an operational setting, luck of the draw can unfairly affect assessment outcomes. As will be shown, MFRM provides a rich set of highly efficient tools to account, and compensate, for rater-dependent measurement error.

The book is organized as follows. In the next chapter, Chapter 2, I briefly describe the principles of Rasch measurement and discuss implications of choosing a Rasch modeling approach to the analysis of many-facet data. Chapter 3 deals with the challenge that rater-mediated assessment poses to assuring high-quality ratings. In particular, I probe into the issue of rater error. The traditional or standard approach to dealing with rater error is to train raters, to compute an index of interrater reliability, and to show that the agreement among raters is sufficiently high. However, in many instances this approach is strongly limited. In order to discuss some of the possible shortcomings and pitfalls, I draw on a sample data set taken from a live assessment of foreign-language writing performance. For the purposes of widening the perspective, I go on describing a conceptual–psychometric framework incorporating multiple kinds of facets that potentially have an impact on the process of rating examinee performance on writing tasks.

In keeping with Step 1 outlined above (Figure 1.1), the potentially relevant facets need to be identified first. Incorporating these facets into a many-facet Rasch measurement (MFRM) model will allow the researcher to closely examine each of the facets and their interrelationships (Step 2). To illustrate the application of such a model (Step 3), I draw again on the writing data and show how that model can be used to gain insight into the many-facet nature of the data (Chapter 4). In Chapters 5 and 6, I pay particular attention to the rater and examinee facets, respectively. In Chapter 7, the discussion focuses on the way raters use the scoring criteria and the different categories of the rating scale.

Chapter 8 illustrates the versatility of the MFRM approach by presenting a number of more advanced models that can be used for analyzing multiple kinds of data and for studying various interactions between facets. The chapter closes with a summary presentation of commonly used models suitable for evaluating the psychometric quality of many-facet data.

Chapter 9 addresses special issues of some practical concern, such as choosing an appropriate rating design, providing informative feedback to raters, and using many-facet Rasch measurement for standard-setting purposes. On a more theoretical note, I deal with differences between MFRM modeling and generalizability theory (G-theory), a psychometric approach rooted in classical test theory that takes different sources of measurement error into account. Finally, I briefly discuss computer programs currently available for conducting a many-facet Rasch analysis, including some extensions of the MFRM approach.

The last chapter, Chapter 10, first provides a summary of major steps and procedures of a standard many-facet Rasch analysis and then presents illustrative MFRM studies drawn from wide-ranging fields of application. After discussing the relationship between measurement and issues of validating performance assessments more generally, the focus shifts toward the potential contribution of MFRM to investigations of rater cognition issues building on mixed methods research designs.