# 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.

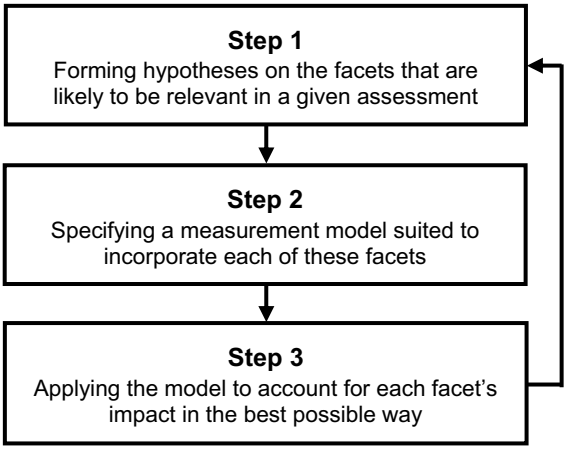


FIG. 1.1 Basic three-step measurement approach to the analysis and evaluation of performance assessments

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.

## 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-faceted* 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. Thus, raters have always played an important role in assessing language proficiency, particularly with respect to the productive skills of writing and speaking. Since the "communicative turn" in language testing, starting around the early 1980s (see, e.g., Bachman, 2000; McNamara, 1996), their role has become even more pronounced. Yet, at the same time, evidence has accumulated pointing to substantial degrees of systematic error in rater judgments that, if left unexplained, may lead to false, inappropriate, or unfair conclusions. For example, it has often been observed that some raters consistently award higher scores than other raters; when these raters are assigned to evaluate the performance of examinees, luck of the draw can unfairly affect assessment outcomes. As will be shown, the MFRM approach provides a rich set of highly efficient tools to account, and compensate, for measurement error, in particular rater-dependent measurement error.

The book is organized as follows. Chapter 2 briefly describes the principles of Rasch measurement and discusses 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 systematic rater error, or rater variability. The traditional or standard approach to dealing with rater error in the context of performance assessments is to train raters in order to achieve a common understanding of what is being measured, 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 proficiency. For the purposes of widening the perspective, I go on describing a conceptual–psychometric framework

incorporating multiple kinds of factors that potentially have an impact on the process of rating examinee performance on writing tasks.

In keeping with Step 1 outlined above, each of the factors and their interrelationships included in the framework constitute a hypothesis about the relevant facets and their influence on the ratings. These hypotheses need to be spelled out clearly and then translated into a many-facet Rasch measurement (MFRM) model in order to allow the researcher to examine each of the hypotheses in due detail (Step 2). To illustrate the application of such a model (Step 3), I draw again on the writing data, study examinees, raters, and criteria as separate facets, and show how that model can be used to gain insight into the many-facet nature of the data (Chapter 4). In doing so, I successively introduce relevant statistical indicators computed in the process of analyzing each of the facets involved, paying particular attention to the rater and examinee facets (Chapters 5 and 6). In Chapter 7, the discussion focuses on the way raters make use of the scoring criteria and the categories of the rating scale.

Chapter 8 illustrates the versatility of the MFRM modeling approach by presenting a number of model variants suited for studying different kinds of data and different combinations of facets. In particular, I look at instantiations of the model addressing constant and variable rating scale structures, and at ways to examine interactions between facets. The chapter closes with a summary presentation of commonly-used model variations suitable for evaluating the psychometric quality of many-facet data.

The last chapter (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 dwell on differences between the MFRM modeling approach and generalizability theory (G-theory), an alternative psychometric approach closely related to classical test theory. Finally, I briefly discuss computer programs currently available for conducting a many-facet Rasch analysis.