Reference Supplement

to the

Manual for

Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment

Section H: Many-Facet Rasch Measurement

Language Policy Division, Strasbourg
October 2009

www.coe.int/lang

Contents

Please cite as follows:

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

Section H

Many-Facet Rasch Measurement

Thomas Eckes
TestDaF Institute, Hagen, Germany

This chapter provides an introductory overview of many-facet Rasch measurement (MFRM). Broadly speaking, MFRM refers to a class of measurement models that extend the basic Rasch model by incorporating more variables (or facets) than the two that are typically included in a test (i.e., examinees and items), such as raters, scoring criteria, and tasks. Throughout the chapter, a sample of rating data taken from a writing performance assessment is used to illustrate the rationale of the MFRM approach and to describe the general methodological steps typically involved. These steps refer to identifying facets that are likely to be relevant in a particular assessment context, specifying a measurement model that is suited to incorporate each of these facets, and applying the model in order to account for each facet in the best possible way. The chapter focuses on the rater facet and on ways to deal with the perennial problem of rater variability. More specifically, the MFRM analysis of the sample data shows how to measure the severity (or leniency) of raters, to assess the degree of rater consistency, to correct examinee scores for rater severity differences, to examine the functioning of the rating scale, and to detect potential interactions between facets. Relevant statistical indicators are successively introduced as the sample data analysis proceeds. The final section deals with issues concerning the choice of an appropriate rating design to achieve the necessary connectedness in the data, the provision of feedback to raters, and applications of the MFRM approach to standard-setting procedures.

The field of language testing draws on a large and diverse set of procedures that aim at assessing a person's language proficiency or some aspect of that proficiency. 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 it is 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. 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.

 Correspondence concerning this chapter or the reproduction or translation of all or part of it should be sent to the author at the following address: Thomas Eckes, TestDaF Institute, Feithstr. 188, 58084 Hagen, Germany. E-mail: thomas.eckes@testdaf.de

1. Facets of Measurement

The first example, the reading comprehension test, describes a frequently encountered measurement situation involving two relevant 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 of an examinee interacts with the difficulty of an item to produce an observed response (i.e., a response to a multiple-choice item scored either correct or incorrect).

The second example, the essay writing, is typical of a situation called *rater-mediated assessment* (Engelhard, 2002; McNamara, 2000). In this kind of situation, one more facet is added to the set of factors that possibly have an impact on examinee scores (besides the examinee and task facets)—the *rater facet*. As we will see 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 may be drawn from the assessment outcomes.

The last example, the face-to-face interview, represents a situation of significantly heightened complexity. At least five facets, and 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, items, or tasks), as well as facets that are assumed to contribute systematic measurement error (e.g., raters, 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 where examinees' career or study plans critically depend on test outcomes. As pointed out previously, factors other than those associated with the construct being measured may have a strong impact on the outcomes of assessment procedures. Therefore, the construction of reliable, valid, and fair measures of language proficiency hinges 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 following three basic steps: *Step 1:* Building hypotheses on which facets are likely to be relevant in a particular testing context. *Step 2:* Specifying a measurement model that is suited to incorporate each of these facets. *Step 3:* Applying the model in order to account for each facet in the best possible way. These steps form the methodological core of a measurement approach to the analysis and evaluation of many-facet data.

2. Purpose and Plan of the Chapter

In this chapter, I present an approach to the measurement of language 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 chapter is on the rater facet and its various ramifications. 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, lenient raters tend to award higher scores than severe raters, and, thus, luck of the draw can unfairly affect assessment outcomes. As will be shown, the MFRM approach provides a rich set of highly flexible tools to account, and compensate, for measurement error, in particular rater-dependent measurement error.

I proceed as follows. In Section 3 below, I briefly look at the implications of choosing a Rasch modeling approach to the analysis of many-facet data. Then, in Section 4, 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 the construct 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 an 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 a writing task.

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 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, specify 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 (Section 5). In doing so, I successively introduce relevant statistical indicators related to the analysis of each of the facets involved, paying particular attention to the rater and examinee facets.

Subsequently, I illustrate 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 (Section 6). In particular, I look at rating scale and partial credit instantiations of the model and at ways to examine interactions between facets. The section closes with a summary presentation of commonly-used model variations suitable for evaluating the psychometric quality of many-facet data. In the last section (Section 7), I address special issues of some practical concern, such as choosing an appropriate rating design, providing feedback to raters, and using many-facet Rasch measurement for standard-setting purposes. Finally, I briefly discuss computer programs currently available for conducting a many-facet Rasch analysis.

3. Rasch Modeling of Many-Facet Data

Many-facet Rasch measurement refers to the application of a class of measurement models that aim at providing a fine-grained analysis of multiple variables potentially having an impact on test or assessment outcomes. MFRM models, or *facets models*, extend the basic *Rasch model* (Rasch, 1960/1980; Wright & Stone, 1979) to incorporate more variables (or facets) than the two that are typically included in a paper-and-pencil testing situation, that is, examinees and items. Facets models belong to a growing family of Rasch models, including the *rating scale model* (RSM; Andrich, 1978), the *partial credit model* (PCM; Masters, 1982), the *linear logistic test model* (LLTM; Fischer, 1973, 1995b; Kubinger, 2009), the *mixed Rasch model* (Rost, 1990, 2004), and many others (for a detailed discussion, see Fischer, 2007; see also Rost, 2001; Wright & Mok, 2004).[1]

1 Early proposals to extend the basic Rasch model by simultaneously taking into account three or more facets ("experimental factors") were made by Micko (1969, 1970) and Kempf (1972). Note also that Linacre's (1989) many-facet Rasch model can be considered a special case of Fischer's (1973) LLTM (see, e.g., Rost & Langeheine, 1997).

Rasch models have a number of distinct advantages over related psychometric approaches that have been proposed in an item response theory (IRT) framework. The most important advantage refers to what has variously been called *measurement invariance* or *specific objectivity* (Bond & Fox, 2007; Engelhard, 2008a; Fischer, 1995a): When a given set of observations shows sufficient fit to a particular Rasch model, examinee measures are invariant across different sets of items or tasks or raters (i.e., examinee measures are "test-free"), and item, task, or rater measures are invariant across different groups of examinees (i.e., item, task, or rater measures are "sample-free").

Measurement invariance implies the following: (a) test scores are *sufficient statistics* for the estimation of examinee measures, that is, the total number correct score of an examinee contains all the information required for the estimation of that examinee's measure from a given set of observations, and (b) the test is *unidimensional*, that is, all items on the test measure the same latent variable or construct. Note that IRT models like the *two-parameter logistic* (2PL) model (incorporating item difficulty and item discrimination parameters) or the *three-parameter logistic* (3PL) model (incorporating a guessing parameter in addition to item difficulty and discrimination parameters) do not belong to the family of Rasch models. Accordingly, they lack the property of measurement invariance (see Kubinger, 2005; Wright, 1999).

Since its first comprehensive theoretical statement (Linacre, 1989), the MFRM approach has been used in a steadily increasing number of substantive applications in the fields of language testing, educational and psychological measurement, health sciences, and others (see, e.g., Bond & Fox, 2007; Engelhard, 2002; Harasym, Woloschuk & Cunning, 2008; McNamara, 1996; Wolfe & Dobria, 2008). As a prominent example, MFRM has formed the methodological cornerstone of the descriptor scales advanced by the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001; see also North, 2000, 2008; North & Jones, 2009; North & Schneider, 1998). In addition, the MFRM approach has been crucial in providing DVDs of illustrative CEFR samples of spoken production for English, French, German, and Italian (see www.coe.int/portfolio; see also Breton, Lepage & North, 2008). Thus, as North (2000, p. 349) put it, many-facet Rasch measurement has been "uniquely relevant to the development of a common framework".

4. Rater-Mediated Performance Assessment

Performance assessments typically employ *constructed-response items*. Such items require examinees to create a response, rather than choose the correct answer from alternatives given. To arrive at scores capturing the intended proficiency, raters have to closely attend to, interpret, and evaluate the responses that examinees provide. The process of performance assessment can thus be described as a complex and indirect one: Examinees respond to test items or tasks designed to represent the underlying construct (e.g., writing proficiency), and raters judge the quality of the responses building on their understanding of that construct, making use of a more or less detailed scoring rubric (Bejar, Williamson & Mislevy, 2006; Freedman & Calfee, 1983; Lumley, 2005; McNamara, 1996; Wolfe, 1997). This long, and possibly fragile, interpretation–evaluation–scoring chain highlights the need to carefully investigate the psychometric quality of rater-mediated assessments. One of the major difficulties facing the researcher, and the practitioner alike, is the occurrence of rater variability.

4.1 Rater variability

The term *rater variability* generally refers to variability that is associated with characteristics of the raters and not with the performance of examinees. Put differently, rater variability is a component of unwanted variability contributing to construct-irrelevant variance in examinee scores. This kind of variability obscures the construct being measured and, therefore, threatens the validity and fairness of performance assessments (Lane & Stone, 2006; McNamara & Roever, 2006; Messick, 1989; Weir, 2005). Related terms like *rater effects* (Myford & Wolfe, 2003, 2004; Wolfe, 2004), *rater error* (Saal, Downey & Lahey, 1980), or *rater bias* (Hoyt, 2000; Johnson, Penny & Gordon, 2009), each touch on aspects of the fundamental rater variability problem.

Rater effects often discussed in the literature are severity, halo, and central tendency effects. The most prevalent effect is the *severity effect*. This effect occurs when raters provide ratings that are

consistently either too harsh or too lenient, as compared to other raters or to established benchmark ratings. As we will see later, severity effects can be explicitly modeled in a MFRM framework. A *central tendency effect* is exhibited when raters avoid the extreme categories of a rating scale and prefer categories near the scale midpoint instead. Ratings based on an analytic rating scheme may be susceptible to a *halo effect*. This effect manifests itself when raters fail to distinguish between conceptually distinct features of examinee performance, but rather provide highly similar ratings across those features; for example, ratings may be influenced by an overall impression of a given performance or by a single feature viewed as highly important. In a MFRM framework, central tendency and halo effects can be examined indirectly (see, e.g., Engelhard, 2002; Knoch, 2009; Linacre, 2008; Myford & Wolfe, 2003, 2004; Wolfe, 2004).

Obviously, then, rater variability is not a unitary phenomenon, but can manifest itself in various forms that each call for close scrutiny. Research has shown that raters may differ not only in the degree of severity or leniency exhibited when scoring examinee performance, but also in the degree to which they comply with the scoring rubric, in the way they interpret and use criteria in operational scoring sessions, in the understanding and use of rating scale categories, or in the degree to which their ratings are consistent across examinees, scoring criteria, performance tasks, testing time, and other facets involved (see Bachman and Palmer, 1996; Brown, 2005; Hamp-Lyons, 2007; Lumley, 2005; McNamara, 1996; Weigle, 2002).

The usual, or standard, approach to come to grips with rater variability, especially in high-stakes tests, consists of three components: rater training, independent ratings of the same performance by two or more raters (repeated ratings), and establishing interrater reliability. The first component, rater training, typically aims at familiarizing raters with the test format, the test tasks, and the rating criteria. More specifically, raters are trained to achieve a *common understanding* of (a) the construct being measured, (b) the level, or levels, of performance the test is aiming at, (c) the criteria and the associated descriptors that represent the construct at each performance level, (d) the categories of the rating scale or scales, and (e) the overall difficulty level of the items or tasks to which examinees are to respond.

Another time-honored safeguard against the occurrence of rater effects is the use of repeated or multiple ratings of the same performance. However, such ratings typically reveal considerable disagreement among raters. In cases of disagreement, those who supervise the raters must decide how to handle the disagreements in order to arrive at a final score. The literature is replete with different procedures that have been proposed to accomplish this, including averaging the complete set of independent ratings, using only those ratings that are in sufficiently close agreement, or calling in a more experienced rater, for example, employing a third-rater adjudication procedure (for a detailed discussion, see Myford & Wolfe, 2002).

Ideally, differences between raters that may still exist after training should be so small as to be practically unimportant; that is, interrater reliability should be as high as possible.[2] Yet, research has shown that this ideal is extremely difficult to achieve in most situations. Raters typically remain far from functioning interchangeably even after extensive training sessions (Barrett, 2001; Eckes, 2004, 2005b; Elbow & Yancey, 1994; Hoyt & Kerns, 1999; Kondo-Brown, 2002; Lumley & McNamara, 1995; O'Sullivan & Rignall, 2007; Weigle, 1998, 1999), and the provision of individualized feedback to raters does not seem to have a sweeping effect either (Elder, Knoch, Barkhuizen & von Randow, 2005; Elder, Barkhuizen, Knoch & von Randow, 2007; Knoch, Read & von Randow, 2007). Moreover, trained, experienced raters have been shown to differ systematically in their interpretation of routinely-used scoring criteria. Rather than forming a single, homogeneous group having a common understanding of how to interpret and use criteria, raters fell into rater types, with each type characterized by a distinct scoring focus. For example, some raters showed a strong focus on criteria referring to vocabulary and syntax, whereas others put significantly more weight on correctness or fluency (Eckes, 2008b, 2009).

Taken together, much more is going on in rater-mediated assessments than can be dealt with satisfactorily in rater training sessions or by computing some index of interrater reliability. Let us take a closer look at this issue.

---

2 Trying to maximize interrater reliability may actually lead to lowering the validity of the ratings, as would be the case, for example, when raters settled for attending to superficial features of examinee performance (see Hamp-Lyons, 2007; Reed & Cohen, 2001; Shohamy, 1995). This clearly unwanted effect is reminiscent of the attenuation paradox in classical test theory (Linacre, 1996; Loevinger, 1954).

4.2 Interrater reliability

4.2.1 The standard approach: Establishing consensus and consistency

As explained above, the trilogy of rater training, repeated ratings, and demonstration of high interrater reliability is the hallmark of the standard approach to solving the rater variability problem. The common, and often undisputed, assumption is that if interrater reliability is sufficiently high, then raters can be said to share the same view of the construct in question and, as a result, will be able to provide accurate ratings in terms of coming close to an examinee's "true score". However, even if high interrater reliability has been achieved in a given assessment context exactly what such a finding stands for may be far from clear. One reason for this is that those reporting on rater performance do not share a common definition of interrater reliability. Over time, interrater reliability has come to be conceptualized in many different ways, by many different people, and for many different purposes, resulting in a bewildering array of indices (see, e.g., Bramley, 2007; Hayes & Krippendorff, 2007; LeBreton & Senter, 2008; Shoukri, 2004; von Eye & Mun, 2005; Zegers, 1991). To complicate matters, different coefficients of interrater reliability can mean vastly different things.

In this situation, it seems reasonable to distinguish between two broad classes of indices: consensus indices and consistency indices (Stemler & Tsai, 2008; Tinsley & Weiss, 1975, 2000). Specifically, a *consensus index* of interrater reliability (also called *interrater agreement*) refers to the extent to which independent raters provide the same rating of a particular person or object (absolute correspondence of ratings). In contrast, a *consistency index* of interrater reliability refers to the extent to which independent raters provide the same relative ordering or ranking of the persons or objects being rated (relative correspondence of ratings).

Though often used interchangeably in the literature, indices from these two classes can lead to discrepant, sometimes even contradictory results and conclusions. It is possible to observe low interrater consensus and, at the same time, high interrater consistency (and vice versa). For example, one rater may award scores to examinees that are consistently one or two scale points lower than the scores that another rater awards to the same examinees. The relative ordering of the examinees will be much the same for both raters, yielding high consistency estimates; yet, the raters have *not* reached exact agreement in any one case.[3]

In the next section, I briefly describe a data set based on a writing proficiency test suited to illustrate the distinction between consensus and consistency indices, as well as some limitations of these indices. Also, I will use the data again in later sections, where I present an illustrative application of the MFRM approach and highlight the advantages of adopting a Rasch measurement perspective.

4.2.2 Sample data: Writing performance assessment

The data considered here concerned examinee performance on the writing section of the Test of German as a Foreign Language (*Test Deutsch als Fremdsprache*, TestDaF).[4] The writing section comprised a single task designed to assess an examinee's ability to produce a coherent and well-structured text on a given topic taken from the academic context. Eighteen raters scored essays written by 307 examinees. Raters were all specialists in the field of German as a foreign language, and they were trained and monitored as to compliance with TestDaF scoring guidelines.

Each essay was rated independently by two raters. In addition, one rater provided ratings of two essays that were randomly selected from each of the other 17 raters' workload. These third ratings ensured that all 18 raters could be directly compared with respect to their severity measures resulting from the MFRM analysis. That is, the additional ratings served to satisfy the basic requirement of a *connected* data set, where all elements are directly or indirectly linked to each other. I will take up this issue in a later section (see Section 7.1 on rating designs).

Ratings were provided on a four-category rating scale, with categories labeled by so-called "TDN levels" (*TestDaF-Niveaustufen*, TestDaF levels, or TDNs, for short). These levels were as follows:

3 There are indices of interrater reliability that belong to both classes. For example, some variants of the intraclass correlation coefficient are a function of both rater consensus and rater consistency (see LeBreton & Senter, 2008; McGraw & Wong, 1996).

4 The TestDaF was administered worldwide for the first time in April 2001. The live examination considered here took place in October 2001 (see, for more detail on the TestDaF performance assessment, Eckes, 2005b, 2008a; see also www.testdaf.de).

below TDN 3, *TDN 3*, *TDN 4*, and *TDN 5*. TDN levels 3 to 5 cover the Council of Europe's (2001) Lower Vantage Level (B2.1) to Higher Effective Operational Proficiency (C1.2).

Raters scored each essay referring to sets of ordered performance descriptors representing three criteria. The first criterion referred to *global impression* (a holistic criterion), the other two criteria were more of an analytic kind, referring to distinct aspects of *treatment of the task* (e.g., completeness) and *linguistic realization* (e.g., vocabulary), respectively. By averaging across the criterion ratings and rounding the obtained averages, final TDNs were awarded to examinees.[5]

4.2.3 Consensus and consistency in essay ratings

Based on the final TDN levels, two indices of interrater consensus (i.e., exact agreement, Cohen's weighted kappa) and two indices of interrater consistency (i.e., product–moment correlation, Kendall's tau-b) were computed. Exact agreement was defined as the number of essays that received identical ratings, divided by the total number of essays rated by the two raters. Kappa corrects the agreement between raters for agreement expected on the basis of chance alone. In the present application, scale categories (i.e., TDN levels) were ordered. Therefore, the weighted version of kappa (Cohen, 1968) based on a linear weighting scheme was used; that is, successively less weight was assigned to disagreement when categories were further apart. Weighted Kappa has a maximum of 1 when agreement is perfect, a value of 0 indicates no agreement better than chance, and negative values show worse than chance agreement (see, e.g., Fleiss, Levin & Paik, 2003; Mun, 2005). Consistency indices computed for the sample data were the product–moment correlation coefficient (also called Pearson's *r*), which reflects the degree of linear relationship between two raters' ratings, and Kendall's tau-b, which reflects the degree of correspondence between two rank orderings of essays, taking tied ranks into account. Both indices take on values between –1 and 1, with higher values indicating a stronger correlation or correspondence, respectively.

Table 1 gives the consensus and consistency results for 14 pairs of raters. The rater pairs are ordered from high to low exact agreement. All raters listed in the table belonged to the panel of 17 operational raters involved in rating examinee writing performance. As mentioned previously, there was one rater (i.e., Rater 06) whose ratings solely served to satisfy the connectedness requirement. Therefore, this rater was not included in the table. The number of common ratings per rater pair varied between 19 essays (rater pair 17/11) and 28 essays (rater pair 15/07).

As can be seen, exact agreement ranged from an acceptably high value of .70 for Raters 07 and 10 to a strikingly low value of .10 for Raters 01 and 14. Most agreement values were in the .40s and .50s, much too low to be satisfactory. Note that weighted kappa reached values that could be judged as sufficiently high only for two pairs (i.e., rater pairs 07/10 and 13/16). In one case, the agreement rate was exactly at a level predicted by chance alone (rater pair 01/14).

Consensus and consistency indices suggested much the same conclusions for the majority of rater pairs. There were two notable exceptions, however. These exceptions concerned Raters 13 and 03, and Raters 05 and 07, respectively. For these two rater pairs, consistency values were moderately high, but consensus values turned out to be much too low to be considered acceptable.

4.2.4 Limitations of the standard approach

To gain insight into the problems associated with the standard approach, first look at rater pair 13/16. For these two raters, fairly good consensus and consistency values were obtained. Table 2 presents the cross-classification of the observed rating frequencies.[6]

Raters 13 and 16 arrived at identical ratings in 12 cases (shown in the shaded cells), they disagreed in eight cases. Each disagreement concerned only one TDN level. For example, four examinees received *TDN 4* by Rater 13, but *TDN 3* by Rater 16. Now look at Rater 13 again, but this time in relation to Rater 03 (see Table 3).

---

5 The rounding rule used here was as follows: average scores smaller than 2.50 were assigned to level *below TDN 3*, average scores from 2.50 to 3.49 to *TDN 3*, average scores from 3.50 to 4.49 to *TDN 4*, and average scores greater than 4.49 to *TDN 5* (for more detail on the procedure of level assignment, see Kecker & Eckes, in press). For purposes of computation, *below TDN 3* was scored "2", the other levels were scored from "3" to "5".

6 In the CEFR Manual (Council of Europe, 2009), tables like these are called "bivariate decision tables".

Table 1. Consensus and Consistency Indices of Interrater Reliability (Sample Data)

		Consensus Indices		*Consistency Indices*	
Rater Pair	*N*	*Exact Agreement*	*Cohen's Weighted Kappa*	*Pearson's r*	*Kendall's Tau-b*
07 / 10	20	.70	.67	.83	.78
13 / 16	20	.60	.67	.84	.84
12 / 03	20	.55	.29	.49	.42
17 / 11	19	.53	.42	.62	.58
14 / 08	23	.52	.50	.77	.70
08 / 12	24	.50	.54	.71	.64
09 / 17	26	.50	.34	.53	.49
05 / 18	21	.48	.53	.76	.68
02 / 04	24	.46	.33	.58	.52
10 / 09	21	.43	.41	.78	.72
15 / 07	28	.36	.20	.53	.48
13 / 03	21	.24	.22	.66	.62
05 / 07	20	.20	.22	.77	.72
01 / 14	20	.10	.00	.21	.26

Note. *N* = number of essays rated. Each essay was independently rated by two trained raters on a four-category rating scale.

There were 16 cases of disagreement, 10 of which concerned one TDN level, and the remaining six cases each concerned two TDN levels. For example, four examinees received *TDN 3* by Rater 13, but *TDN 5* by Rater 03. However, the disagreements appeared to be anything but random. There was not a single case in which Rater 03 provided a lower rating than Rater 13. Thus, Rater 03 exhibited a tendency to award *systematically higher* levels than Rater 13.

The pattern of disagreements for pair 13/03 suggests the following tentative conclusion: Rater 13 disagreed with Rater 03 so strongly because he or she was more *severe* than Rater 03, or, conversely, because Rater 03 was more *lenient* than Rater 13.

This difference in severity or leniency, respectively, could account for the fact that consensus indices were unacceptably low, whereas consistency indices were considerably higher. As explained previously, consistency indices of interrater reliability are sensitive to the relative ordering of examinees. These orderings of examinees, as evident in each of the raters' TDN level assignments, were indeed highly congruent.

Given that this conclusion is correct: What about the high reliability indices (in terms of both consensus and consistency) observed for Raters 13 and 16? Could it be that these two raters were characterized by much the same degree of severity or leniency, respectively, and that on these grounds they provided highly similar ratings in the majority of cases? And, when similar degrees of severity/leniency accounted for satisfactorily high consensus and consistency observed for these two raters, would it be reasonable to call their ratings "accurate"?

These questions point to a fundamental problem of the standard approach to interrater reliability, a problem that may be dubbed the *agreement–accuracy paradox*. High consensus or agreement among raters, and in this sense, high reliability, does not necessarily imply high accuracy in assessing examinee proficiency. Neither does high consistency imply high accuracy, even if consensus is high. Thus, high reliability may lead to the wrong conclusion that raters provided highly accurate ratings when in fact they did not.

Now, what about raters showing low consensus *and* low consistency? One may be tempted to conclude, as quite a number of researchers have done, that their ratings are useless and that they should be excluded from the panel of raters, replacing them by others who show much higher reliability (see, e.g., Tinsley & Weiss, 1975, 2000). Let us look at a final example illuminating this point (see Table 4).

Table 2. Cross-Classification of Rating Frequencies for Raters 13 and 16

	Rater 16				
Rater 13	*b. TDN 3*	*TDN 3*	*TDN 4*	*TDN 5*	*Row total*
below TDN 3	**8**				8
TDN 3	**1**	**1**			2
TDN 4		**4**	**2**	**3**	9
TDN 5				**1**	1
Column total	9	5	2	4	20

Note. Consensus indices are .60 (exact agreement) and .67 (Cohen's weighted kappa). Consistency indices are .84 (Pearson's *r*) and .84 (Kendall's tau-b).

Table 3. Cross-Classification of Rating Frequencies for Raters 13 and 03

	Rater 03				
Rater 13	*b. TDN 3*	*TDN 3*	*TDN 4*	*TDN 5*	*Row total*
below TDN 3	**1**	**3**	**2**		6
TDN 3			**5**	**4**	9
TDN 4			**2**	**2**	4
TDN 5				**2**	2
Column total	1	3	9	8	21

Note. Consensus indices are .24 (exact agreement) and .22 (Cohen's weighted kappa). Consistency indices are .66 (Pearson's *r*) and .62 (Kendall's tau-b).

Table 4 presents the cross-classification of rating frequencies for Raters 01 and 14. Of all raters considered in the present sample, these two raters had the lowest consensus and consistency values. They agreed exactly in only two out of 20 cases. Note, however, that the distribution of rating frequencies still seemed to bear some regularity. Specifically, both raters used a restricted range of the TDN scale, that is, Rater 14 only used scale categories *TDN 3* to *TDN 5*, and Rater 01 only used the two highest scale categories (i.e., *TDN 4* and *TDN 5*). Moreover, Rater 01 awarded higher scores in 17 cases, suggesting a tendency to be more lenient than the other rater. Clearly, this kind of regularity is missed when the standard approach is adopted.

Employing alternative consensus or consistency indices of interrater reliability is no way out of this dilemma. The basic difficulties in adequately representing the structure inherent in the rating data will remain unchanged as long as the underlying rationale is the same. The paradox exemplified here can only be resolved when the standard approach is abandoned in favor of a measurement approach.

Table 4. Cross-Classification of Rating Frequencies for Raters 01 and 14

	Rater 14				
Rater 01	*b. TDN 3*	*TDN 3*	*TDN 4*	*TDN 5*	*Row total*
below TDN 3					0
TDN 3					0
TDN 4		**6**	**1**	**1**	8
TDN 5		**5**	**6**	**1**	12
Column total	0	11	7	2	20

Note. Consensus indices are .10 (exact agreement) and .00 (Cohen's weighted kappa). Consistency indices are .21 (Pearson's *r*) and .26 (Kendall's tau-b).

Many-facet Rasch measurement yields a detailed analysis of the similarities and differences in raters' views when assessing examinees' language proficiency. In a later section, I demonstrate how this issue can be dealt with using a MFRM approach. First, however, I want to broaden the perspective and go into somewhat more detail regarding the various sources of variability in ratings that are typical of writing performance assessments.

4.3 A conceptual–psychometric framework

The MFRM analysis of the sample performance data rests on a conceptual model of factors that typically influence ratings of examinee writing performance. Figure 1 depicts these factors and their mutual relationships (see Eckes, 2005a, 2008a).

To be sure, the factors shown do not encompass all that may happen in a particular rating session. The rating process is undoubtedly far more complex and dynamic than can be summarized in a diagram, and the factors coming into play are diverse at any given moment (see, e.g., Engelhard & Myford, 2003; Lane & Stone, 2006; Murphy & Cleveland, 1995).

Each of the factors, as well as each of the factor interrelations, deemed important in a particular context constitutes a hypothesis about the potential sources of variation in the ratings. These hypotheses may originate from previous research on the subject matter, from observations made in the particular kind of assessment setting, or from earlier modeling attempts that turned out to be insufficient or incomplete. In any case, failing to identify relevant facets can produce misleading measurement results. For example, unidentified or "hidden" facets may yield biased estimates of examinee proficiency or rater severity.

Note also that the diagram refers to factors usually involved in *writing* performance assessments. Assessing *speaking* performance is often more intricate still, particularly in direct speaking tests (Berry, 2007; Brown, 2005; Fulcher, 2003; O'Sullivan, 2008). For example, when speaking proficiency is assessed through face-to-face interaction, interviewers/interlocutors and other examinees simultaneously present in the assessment situation, as in a group oral test (Van Moere, 2006), have to be considered as additional factors affecting examinee performance.

With these caveats in mind, the following outline will help to prepare the stage for introducing more specific concepts relevant for a detailed, psychometric analysis of performance assessments.

4.3.1 Proximal and distal factors

Consider first the factors shown in the middle part of the diagram. This part comprises factors that have an immediate impact on the scores awarded to examinees. The most important of these factors, which may be called *proximal* factors, is of course the construct being measured (i.e., examinee language proficiency).

Other proximal factors are basically irrelevant to the construct and thus potentially contribute to systematic measurement error in the ratings. These include (a) rater effects, in particular severity, central tendency, and halo effects, (b) variability in the difficulty of the tasks presented to examinees, and (c) variability in the difficulty of scoring criteria. Finally, a less obvious source of measurement error concerns the variability in the structure of the rating scale used. That is, the ordered categories of a given rating scale may change their meaning between raters, within raters over time, between tasks or between criteria. For example, raters may differ from each other in their interpretation of the ordering of scale categories; that is, some raters may actually perceive two adjacent categories in terms of the implied performance levels to be much closer together than other raters do.

The left-hand side of Figure 1 shows three categories of *distal* variables that exert additional influence on the ratings, albeit usually in a more indirect and diffuse way: (a) features of examinees (e.g., gender, ethnicity, first language, personality traits, beliefs, goals), (b) features of raters (e.g., number of foreign languages spoken, professional background, educational career, goals and motivation), and (c) features of the situation, that is, features of the assessment or rating context (e.g., technical and physical environment, rater workload, time of rating, quality management policy, organizational values). Some of these distal factors may interact with one another and may also interact with some of the proximal factors, such as when examinee gender interacts with rater severity or when raters' degree of professional experience interacts with their interpretation and use of scoring criteria.

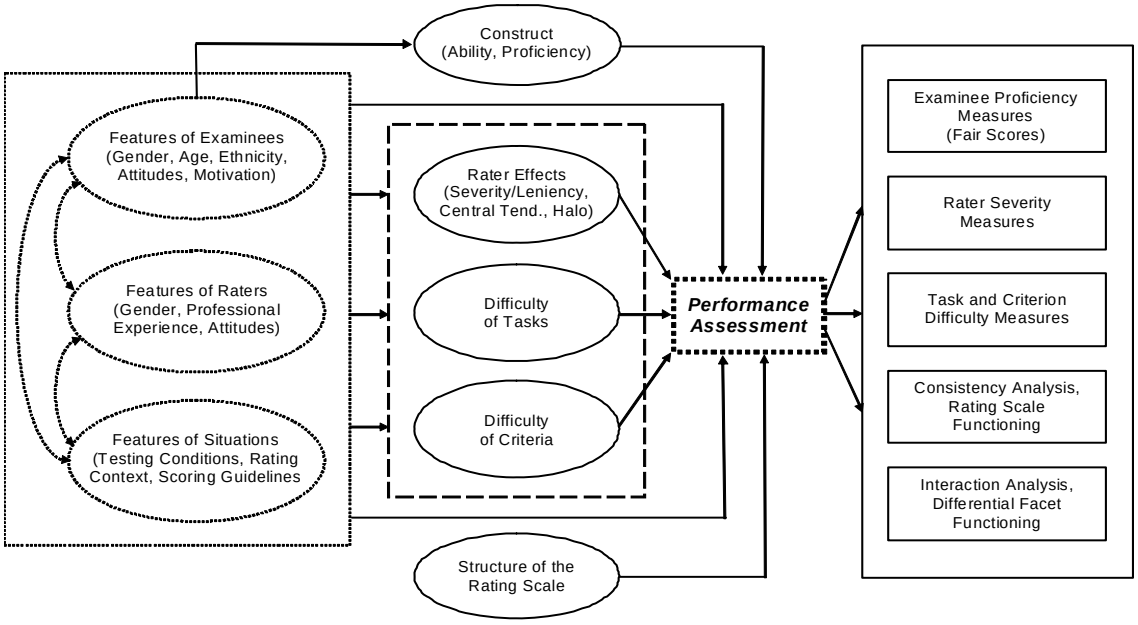

Figure 1. A conceptual–psychometric framework of factors relevant in rater-mediated performance assessments.

4.3.2 Measurement outcomes

On the right-hand side, the diagram lists major types of output from a MFRM analysis of writing performance assessments. MFRM modeling generally provides detailed insight into the functioning of each factor (proximal and/or distal) that is deemed relevant in the particular assessment context. In the following, basic concepts are introduced in a non-technical manner. More detail, including formal definitions of statistical indicators, is provided in later sections.

As mentioned earlier, the MFRM model is an extension of the basic Rasch model. This extension is twofold: (a) there is no restriction to the analysis of only two facets (i.e., examinees and items), and (b) the data being analyzed need not be dichotomous. In an analysis of performance assessments, the MFRM model allows one to take account of additional facets of that setting that may be of particular interest, such as raters, tasks, and criteria. Moreover, raters typically award scores to examinees using ordered scale categories (i.e., rating scales). Therefore, the data is polytomous in most instances (see, e.g., de Ayala, 2009; Embretson & Reise, 2000; Ostini & Nering, 2006).

Within each facet, the model represents each element (i.e., each individual examinee, rater, task, criterion, etc.) by a separate parameter value. The parameters denote distinct attributes of the facets involved, such as proficiency (for examinees), severity (for raters), and difficulty (for items, tasks, or scoring criteria). In many assessment contexts, the measure of primary interest refers to examinees. Specifically, for each examinee, a MFRM analysis provides a *proficiency measure* which is expressed in an equal-interval metric, that is, in *log-odds units* or *logits*. When a set of empirical data fits the model, these measures compensate for rater severity/leniency differences; that is, the examinee proficiency measures are independent of the particular sample of the raters who provided the ratings. In addition, the analysis provides a *standard error* that indicates the precision of each proficiency measure.

On the basis of MFRM model parameter estimates, a *fair score* (fair average, expected score) can be derived for each examinee (Linacre, 2008). Fair scores result from a transformation of examinees' proficiency estimates reported in logits to the corresponding scores on the raw-score scale. That is, a fair score is the score that a particular examinee would have obtained from a rater of average severity. Fair scores thus illustrate the effect of the model-based compensation for rater severity/leniency differences.

When a MFRM analysis is run, the specified facets are analyzed simultaneously and calibrated onto a single linear scale (i.e., the logit scale). The joint calibration of facets makes it possible to measure rater severity on the same scale as examinee proficiency, task difficulty, and criterion difficulty. By placing all parameter estimates on a common scale, a frame of reference for interpreting the results of the analysis is constructed. Therefore, measures of examinee proficiency, rater severity, task difficulty, and criterion difficulty can be directly compared to each other.

A MFRM analysis provides, for each element of each facet, *fit indices* showing the degree to which observed ratings match the expected ratings that are generated by the model. Regarding the rater facet, fit indices provide estimates of the consistency with which each individual rater made use of the scale categories across examinees, tasks, and criteria. A *consistency analysis* based on the inspection of rater fit indices has an important role to play in rater monitoring and rater training, especially when it comes to provide feedback to raters on their rating behavior. Fit indices also help to detect various rater effects besides severity/leniency, such as central tendency or halo effects (Engelhard, 2002; Knoch, Read & von Randow, 2007; Myford & Wolfe, 2003, 2004; Wolfe, 2004).

In performance assessments, the input data to a MFRM analysis are generally ratings provided on an ordinal scale. How well the categories on a particular scale, that is, the scores awarded to examinees, are separated from one another is an empirical question directly relevant to establishing the psychometric quality of the data. A MFRM analysis typically provides a number of useful indices for studying the functioning of rating scales. For example, for each rating scale category, the average of the examinee proficiency measures that went into the calculation of the category calibration measure should advance monotonically with categories. When this pattern is borne out in the data, the results suggest that examinees with higher ratings are indeed exhibiting "more" of the variable that is being measured than examinees with lower ratings.

Once the parameters of a MFRM model have been estimated, possible interaction effects, such as the interaction between raters and examinees or between examinees and tasks, can be investigated. To this end, the basic MFRM model needs to be extended to include interaction terms that represent the deviation of particular combinations of between-facet elements (e.g., rater–examinee pairs) from their average parameter estimates (raters and examinees, respectively). An *interaction analysis* may thus identify unusual interaction patterns among various facet elements, particularly those patterns that suggest consistent deviations from what is expected on the basis of the model. The occurrence of such deviations would indicate the presence of *differential facet functioning* (Du, Wright & Brown, 1996; Engelhard, 2002; Wang, 2000).

5. A Sample Data MFRM Analysis

In this section, I illustrate the application of the MFRM modeling approach. In so doing, I build on the writing performance sample data used previously.

The data was analyzed by means of the computer program FACETS (Version 3.64; Linacre, 2008). FACETS used the scores that raters awarded to examinees on each of the three criteria (i.e., *global impression*, *treatment of the task*, *linguistic realization*) to estimate individual examinee proficiencies, rater severities, criterion difficulties, and scale category difficulties. The program calibrated the examinees, raters, and criteria, as well as the rating scale onto the same equal-interval scale (i.e., the logit scale), creating a single frame of reference for interpreting the results of the analysis.

Modeling details and measurement results are presented in an order that aims to facilitate the understanding of the basic rationale. First, I specify the MFRM model on which most of the analysis was based. A graphical display illustrates the joint calibration of examinees, raters, criteria, and the rating scale. Subsequently, I present detailed measurement results for each facet separately, beginning with the rater facet, followed by results for the examinee and criterion facets. The focus is on assessing the degree of data–model fit for the rater facet as revealed by rater fit statistics, on comparing raters with respect to their severity or leniency, as well as on assessing the degree of variability within this facet as summarized by rater separation statistics. Measurement results for examinees highlight ways to deal with the issue of fair assessment in light of substantial rater variability. Finally, I discuss results concerning the functioning of the rating scale and very briefly look at the issue of global model fit.

5.1 The MFRM model

The many-facet Rasch measurement model used to analyze the writing performance sample data can be specified as follows:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \tau_k, \quad (1)$$

where

p_{nijk} = probability of examinee n receiving a rating of k on criterion i from rater j,
p_{nijk-1} = probability of examinee n receiving a rating of $k-1$ on criterion i from rater j,
θ_n = proficiency of examinee n,
β_i = difficulty of criterion i,
α_j = severity of rater j,
τ_k = difficulty of receiving a rating of k relative to a rating of $k-1$.

The category coefficient, τ_k, is the location where the adjacent categories, k and $k-1$, are equally probable to be observed. In other words, τ_k represents the transition point at which the probability is 50% of an examinee being rated in one of two adjacent categories, given that the examinee is in one of those two categories. These transition points are also called *Rasch-Andrich thresholds* (see Bond & Fox, 2007; Linacre, 2006a; see also Andrich, 1998).

From an applied point of view, the category coefficient, or threshold parameter, indicates how the rating data are to be handled. In Equation 1, the parameter specifies that a rating scale model (Andrich, 1978) should be used; that is, in the analysis, the four-category scale is treated as if all scoring criteria shared the same rating scale structure, with category coefficients calibrated jointly across the three criteria. Hence, Equation 1 is the expression for a *three-facet rating scale model* (Linacre & Wright, 2002). Alternatively, the threshold parameter could be specified in such a way as to allow for variable rating scale structures (see Section 6.3).

As can be seen from Equation 1, a MFRM model is essentially an additive linear model that is based on a logistic transformation of observed ratings to a logit or log-odds scale ("ln" = natural logarithm). The logistic transformation of ratios of successive category probabilities (log odds) can be viewed as the dependent variable with various facets, such as examinees, raters, and criteria conceptualized as independent variables that influence these log odds. Note also that incomplete rating designs (i.e., missing data) are accommodated by this model because it is only evaluated for observed data points. There is no requirement to impute, or adjust for, unobserved data.

In order to establish the origin of the logit scale and make the model identifiable, I centered the rater and criterion facets; that is, these facets were constrained to have a mean element measure of zero. Another identification constraint required that the sum of the category coefficients equaled zero. As usual, the examinee facet was the only facet left non-centered.

5.2 Variable map

Figure 2 displays the variable map representing the calibrations of examinees, raters, criteria, and the four-category TDN rating scale as raters used it to score examinee essays. This map (also called "Wright map") is a very informative piece of output from the analysis, portraying all the facets of the analysis in a single frame of reference and thus facilitating comparisons within and between the various facets.

The logit scale appears as the first column in the map. All measures of examinees, raters, and criteria, as well as the category coefficients, are positioned on this scale.

The second column (labeled "Examinee") displays the estimates of examinee proficiency on the TestDaF writing section. In this example, each star represents three examinees, and a dot represents one

Logit	Examinee	Rater	Criterion	TDN Scale
	High	*Severe*	*Hard*	(TDN 5)
8	.			
7	.			
6	*.			
	*.			
	*.			
5	*			
	.			
	**			
4	***.			
	**.			-----
	**.			
3	**.			
	***			
	***	16		
2	*******.	13		
	***.	14		TDN 4
	***.	09 15		
1	*****	05		
	***.		LR	
	****.	04	TT	
0	***	06 08 11		-----
	*******	18		
	***.	17		
-1	***.	10 12	GI	
	***	02		
	****			
-2	**.	03		TDN 3
	**.	01 07		
	**.			
-3	**.			
	*.			
	**.			-----
-4	.			
	.			
	.			
-5	*			
	.			
	.			
-6	.			
-7	.			
	Low	*Lenient*	*Easy*	(below 3)

Figure 2. Variable map from the many-facet rating scale analysis. Each star in the second column represents three examinees, and a dot represents one or two examinees. Scoring criteria in the fourth column are as follows: LR = linguistic realization, TT = treatment of the task, GI = global impression. The horizontal dashed lines in the rightmost column indicate the category threshold measures.

or two examinees. Proficiency measures are ordered with higher-scoring examinees appearing at the top of the column, and lower-scoring examinees appearing at the bottom. Note that this is a positively-oriented facet, as indicated by the plus sign before the examinee parameter θ_n in Equation 1; that is, the higher the examinee measure, the higher the raw score.[7]

The third column (labeled "Rater") compares the raters in terms of the level of severity or leniency each exercised when rating essays. More severe raters appear higher in the column, while more lenient raters appear lower. Thus, in this analysis, the rater facet has a negative orientation, as indicated by the minus sign before the α_j parameter in Equation 1; that is, the higher the rater measure the lower the raw score. In principle, the measurement model could also be defined in terms of rater leniency, instead of rater severity. Then the rater term would be positive in the measurement model, and the rater column in Figure 2 would be reversed: lenient raters at the top, severe raters at the bottom.

As can be seen, the variability across raters in their level of severity was substantial. In fact, the rater severity measures showed a 4.64-logit spread, which was about a third (31.1%) of the logit spread observed for examinee proficiency measures (14.93 logits). Thus, despite all efforts at achieving high rater agreement during extensive training sessions, the rater severity measures were far from being homogeneous. This striking lack of consensus among raters would have a considerable impact on classification-level decisions.

The fourth column (labeled "Criterion") compares the three scoring criteria in terms of their relative difficulties. Criteria appearing higher in the column were more difficult than those appearing lower. That is, the higher the difficulty measure of a particular criterion, the more difficult it was for examinees to receive a high score on that criterion. The criterion facet, as specified in Equation 1, is negatively oriented. As can be seen, *linguistic realization* and *treatment of the task* were similarly difficult, *global impression* was the easiest one.

The last column maps the four-category TDN scale to the equal-interval logit scale. The lowest scale category (*below TDN 3*) and the highest scale category (*TDN 5*) both of which would indicate extreme ratings, are shown in parentheses only. This is because the boundaries of the two extreme categories are $-\infty$ (for the lowest one) and $+\infty$ (for the highest one). Each horizontal dashed line is positioned at the *category thresholds*, or *Rasch-half-score-point thresholds*, that is, at the locations where the average expected score on the rating scale is "category + 0.5" score points. Put differently, these thresholds define intervals on the latent variable in which the rounded expected scores are the integer category values. For example, category value 4 (representing *TDN 4*) is, on the average, expected for examinees with measures that fall in the interval between –0.06 logits and 3.76 logits (for a discussion of these and related threshold conceptualizations, see Linacre, 2006a).

5.3 Rater measurement results

5.3.1 Rater severity and rater fit

Figure 2 clearly showed that the raters studied here varied widely in their measures of severity. Detailed measurement results on each individual rater are presented in Table 5. The raters are ordered from most severe to most lenient. To the right of each severity measure is the standard error (*SE*), indicating the precision with which that measure was estimated. Other things being equal, the greater the number of ratings an estimate is based on, the smaller its standard error. For example, the severity measure of Rater 07 (–2.24 logits) was estimated with the highest precision (*SE* = 0.15), based on a total of 204 ratings (i.e., 68 essays rated on 3 criteria each); the lowest precision was obtained for the estimate of Rater 16's measure (2.40 logits, *SE* = 0.30), based on 60 ratings (i.e., 20 essays rated on 3 criteria each).

A large number of factors may contribute to a rater's tendency to rate harshly or leniently, such as those referring to professional experience, personality traits, attitudes, demographic characteristics, workload, and assessment purpose. For example, the most experienced or senior rater may also be the most severe. That rater may feel that he or she must "set the standard" for the other raters by noticing

7 For ease of presentation, examinees with extreme scores are not shown here (9 examinees had *below TDN 3* in all criteria, exactly the same number of examinees had *TDN 5* in all criteria); one of these examinees received extreme scores through the third ratings as well. Thus, non-extreme scores were available for 1,833 responses.

Table 5. Measurement Results for the Rater Facet

Rater	*Severity Measure*	*SE*	*Infit*	*Outfit*	*Fair Average*	*Obs. Average*	*Number of Ratings*
16	2.40	0.30	0.93	0.80	3.00	3.03	60
13	2.09	0.20	0.82	0.74	3.08	3.11	123
14	1.83	0.18	1.10	1.09	3.15	3.45	129
15	1.21	0.22	1.39	1.43	3.32	3.58	84
09	1.21	0.17	0.81	0.79	3.32	3.39	141
05	1.05	0.19	1.12	1.06	3.37	3.37	123
04	0.29	0.23	0.89	0.87	3.60	3.72	72
11	0.16	0.26	0.75	0.75	3.63	3.54	57
08	0.14	0.18	1.05	1.07	3.64	3.49	141
06	0.09	0.20	1.11	1.08	3.65	3.59	102
18	–0.17	0.27	1.30	1.39	3.73	3.81	63
17	–0.57	0.18	0.81	0.83	3.83	3.98	135
12	–1.00	0.18	1.08	1.09	3.94	3.61	132
10	–1.02	0.19	1.02	0.99	3.94	3.48	123
02	–1.23	0.24	1.16	1.17	3.99	4.10	72
03	–2.01	0.19	0.82	0.74	4.17	4.02	123
01	–2.23	0.29	0.96	1.23	4.23	4.52	60
07	–2.24	0.15	0.94	0.92	4.23	4.06	204

Note. *SE* = Standard error. Infit and outfit are mean-square statistics.

even small flaws in examinee performance that are otherwise likely to be overlooked. Conversely, less-experienced raters may tend to give the benefit of the doubt to examinees, especially when performances are at the border of two adjacent proficiency levels. There has been a notable lack of research into the personal and situational determinants of rater severity (for steps in this direction, see Eckes, 2008b; McManus, Thompson & Mollon, 2006; Myford, Marr & Linacre, 1996; Stone, 2006; see also Landy & Farr, 1980). Research along these lines would also need to address the issue of stability and change in rater severity (see, e.g., Congdon & McQueen, 2000a; Lamprianou, 2006; Lunz, 2007; O'Neill & Lunz, 2000).

The next two columns of Table 5 present statistical indicators of the degree to which raters used the TDN scale in a *consistent* manner. These indicators are also called rater fit statistics. In the present analysis, *rater fit* refers to the extent to which a given rater is associated with unexpected ratings, summarized over examinees and criteria.

Rater fit statistics can be formally derived as follows. Referring to the model specified in Equation 1, the probability of examinee n receiving a rating of k ($k = 0, \ldots, m$) on criterion i from rater j is

$$p_{nijk} = \frac{\exp\left[k(\theta_n - \beta_i - \alpha_j) - \sum_{s=0}^{k} \tau_s\right]}{\sum_{r=0}^{m} \exp\left[r(\theta_n - \beta_i - \alpha_j) - \sum_{s=0}^{r} \tau_s\right]}, \tag{2}$$

where τ_0 is defined to be 0. The denominator in Equation 2 is a normalizing factor based on the sum of the numerators.

Generally, fit statistics indicate the degree to which observed ratings match the expected ratings that are generated by the MFRM model. Let x_{nij} be the observed rating for examinee n by rater j on criterion i, and e_{nij} be the expected rating, based on Rasch parameter estimates. Differences between observed and expected ratings can then be expressed as standardized residuals:

$$z_{nij} = \frac{x_{nij} - e_{nij}}{w_{nij}^{1/2}}, \tag{3}$$

where

$$e_{nij} = \sum_{k=0}^{m} k p_{nijk} \tag{4}$$

and

$$w_{nij} = \sum_{k=0}^{m} (k - e_{nij})^2 \, p_{nijk}. \tag{5}$$

In Equation 5, w_{nij} is the model variance of the observation around its expectation under Rasch-model conditions.

Large standardized residuals for individual raters may indicate the occurrence of rater inconsistency. Standardized residuals with absolute values greater than 2 have $p < .05$ under Rasch-model conditions, and so indicate significant departure in the data from the Rasch model. Those observations are commonly considered significantly unexpected (Engelhard, 2002; Myford & Wolfe, 2003).

When standardized residuals are squared, and the squared standardized residuals are summarized over different facets and different elements within a facet, indices of data–model fit are obtained. These summary statistics are called *mean-square fit statistics*. They have the form of chi-square statistics divided by their degrees of freedom (R. M. Smith, 2004; Wright & Masters, 1982).

To derive a mean-square fit statistic for rater j, the squared standardized residuals are averaged over all examinees $n = 1, \ldots, N$, and criteria $i = 1, \ldots, I$, rated by that rater:

$$MS_{U(j)} = \frac{\sum_{n=1}^{N} \sum_{i=1}^{I} z_{nij}^2}{N \cdot I}. \tag{6}$$

Equation 6 gives the *unweighted* mean-square fit statistic for rater j. The unweighted fit statistic is also called *outfit*. Rater outfit is particularly sensitive to occasional highly unexpected ratings from an otherwise consistent rater ("outfit" is short for "outlier-sensitive fit statistic").

Less sensitive to outlying unexpected ratings is the *weighted* mean-square fit statistic:

$$MS_{W(j)} = \frac{\sum_{n=1}^{N} \sum_{i=1}^{I} w_{nij} z_{nij}^2}{\sum_{n=1}^{N} \sum_{i=1}^{I} w_{nij}}, \tag{7}$$

where w_{nij} is defined as in Equation 5.

The weighted fit statistic given in Equation 7 is also called *infit*. Rater infit provides an estimate of the consistency with which a particular rater uses the rating scale across examinees and criteria; that is, this statistic is sensitive to an accumulation of unexpected ratings ("infit" is short for "information weighted fit statistic"). For this reason, infit is commonly considered more important than outfit in judging model fit (see, e.g., Linacre, 2002c, 2008; Myford & Wolfe, 2003).

Infit and outfit statistics have an expected value of 1 and can range from 0 to infinity (Linacre, 2002c; Myford & Wolfe, 2003). Raters with fit values greater than 1 show more variation than expected in their ratings; data provided by these raters tend to *misfit* (or *underfit*) the model. By contrast, raters with fit values less than 1 show less variation than expected; data provided by these raters tend to *overfit* the model. Misfit is generally deemed to be more problematic than overfit (Myford & Wolfe, 2003).

As a rule of thumb, Linacre (2002c, 2008) suggested 0.50 as a lower-control limit and 1.50 as an upper-control limit for infit and outfit statistics. That is, Linacre considered mean-square values in the range between 0.50 and 1.50 as "productive for measurement" or as indicative of "useful fit" (see also Linacre, 2003b). Other researchers suggested using a narrower range defined by a lower-control limit of 0.70 (or 0.75) and an upper-control limit of 1.30 (see, e.g., Bond & Fox, 2007; McNamara, 1996; Wright & Linacre, 1994). Su, Sheu and Wang (2007) proposed evaluating infit and outfit mean-square statistics on the basis of confidence intervals estimated through bootstrapping (see also Wolfe, 2008). Generally speaking, the actual definition of lower- and upper-control limits for mean-square fit statistics will depend in part on the nature of the assessment purpose (e.g., high-stakes vs. low-stakes decisions) and on the resources available for studying rater misfit.[8]

As can be seen from Table 5, most raters had mean-square fit statistics that stayed within a narrowly defined fit range. Two raters (Rater 15 and Rater 18) showed a somewhat heightened degree of misfit, whereas Rater 11 exhibited a slight tendency towards overfit.

Rater misfit can indicate an idiosyncratic rating style or otherwise overly inconsistent rating behavior. However, attention must also be paid to possible idiosyncratic examinee performance, which may be reflected in the ratings awarded. Overfitting raters typically provide muted ratings that suggest a central tendency or, alternatively, a halo effect (see Engelhard, 2002; Myford & Wolfe, 2004). Moreover, in paired rating-designs such as the one underlying the present data, overfit can also indicate when raters are colluding. For instance, if two insecure raters are paired together, they may consult in order to be "on the safe side". Also, if there is a penalty for too much disagreement between a pair of raters, one rater may try to imitate the rating style of the paired rater.

5.3.2 Fair average and observed average

The last two columns in Table 5 display statistics that help to gain a substantive interpretation of rater severity differences and their implications: fair average and observed average. Both kinds of averages are in the raw-score metric, that is, in the metric of the TDN scale.

An observed average for rater j, that is, $M_{O(j)}$, is the rater's mean rating across all examinees and criteria that he or she rated:

$$M_{O(j)} = \frac{\sum_{n=1}^{N} \sum_{i=1}^{I} x_{nij}}{N \cdot I}. \tag{8}$$

A non-trivial problem with observed averages is that they confound rater severity and examinee proficiency. For example, when a particular rater's observed average is markedly lower than other raters' observed averages, this could be so because the rater was more severe than the other raters, or because the rater had more examinees of lower proficiency to rate. Fair averages resolve this problem: A fair average for rater j adjusts the observed average $M_{O(j)}$ for the difference in the level of proficiency in j's sample of examinees from the examinee proficiency mean across all raters. Fair averages thus disentangle rater severity from examinee proficiency.

To compute a fair average for rater j, the parameter estimates of all elements of the other facets that participated in producing the observed scores, except for rater j's severity parameter, are set to their mean values. In the present three-facet example, Equation 1 becomes

$$\ln\left[\frac{p_{jk}}{p_{jk-1}}\right] = \theta_M - \beta_M - \alpha_j - \tau_k, \tag{9}$$

8 For the purposes of significance testing, infit and outfit, respectively, can be transformed into a t statistic, or *standardized fit statistic*, that follows approximately the standard normal distribution. According to Linacre (2003b), the standardized fit statistic tests the null hypothesis that the data fit the model "perfectly", whereas the mean-square statistics indicate whether the data fit the model "usefully". Note that, as sample size increases, ever smaller deviations from model expectations (i.e., mean square = 1.0) will become statistically significant.

where p_{jk} is the probability of rater j using category k across all examinees and criteria, and θ_M and β_M are the mean examinee proficiency and the mean criterion difficulty measures, respectively.

The fair average (or expected score) for rater j, that is, $M_{F(j)}$, is then given as:

$$M_{F(j)} = \sum_{r=0}^{m} r p_{jr}. \tag{10}$$

In the present sample data analysis, the rater severity measures, the range of which is theoretically infinite, are transformed back to the raw-score scale, which has a lower bound of 2 (rating category *below TDN 3*) and an upper bound of 5 (rating category *TDN 5*).

Fair averages enable fair comparisons between raters to be made in the raw-score metric. For example, comparing the fair averages of Rater 16 and Rater 07, it would be safe to conclude that, on average, Rater 16 gave ratings that were 1.23 raw-score points lower than Rater 07. That is, the severity difference between these two raters exceeded one TDN level.

5.3.3 Rater separation

The distribution of rater severity measures depicted in the variable map (see Figure 2) pointed to a pronounced between-rater heterogeneity. To summarize observations like these, and to provide a sound basis for drawing conclusions from them, several group-level statistical indicators are available. These so-called *separation statistics* are computed for each facet specified in the model (Myford & Wolfe, 2003; Schumacker & E. V. Smith, 2007; Wright & Masters, 1982). Next, I discuss four particularly useful separation statistics as they relate to the rater facet.

The first statistic, the *homogeneity statistic*, provides a test of the null hypothesis that rater severity measures in the population are all the same, after accounting for measurement error (Hedges & Olkin, 1985; Linacre, 2008). This fixed (all same) statistic is:

$$Q = \sum_{j=1}^{J} w_j (\hat{\alpha}_j - \hat{\alpha}_+)^2, \tag{11}$$

where

$$\hat{\alpha}_+ = \sum_{j=1}^{J} w_j \hat{\alpha}_j \Big/ \sum_{j=1}^{J} w_j \tag{12}$$

and $w_j = 1/SE_j^2$. As before, SE_j is the standard error that is associated with the estimate of the severity parameter for rater j. This estimate is denoted by $\hat{\alpha}_j$.

Q is approximately distributed as a chi-square statistic with $df = J - 1$ (*df* is short for *degrees of freedom*). In practice, a significant Q value for a given sample of raters indicates that the severity measures of at least two of the J raters in the population are different. Note that Q is very sensitive to sample size. Hence, Q may reach the level of significance, particularly in large samples, even though the actual rater severity differences are fairly small. In the present small-sample analysis, where $J = 18$, Q was highly significant ($Q = 1{,}221.8$, $df = 17$, $p < .01$).

When the null hypothesis of equal severity measures has been rejected, the difference in severity measures of any two raters j and k ($j, k = 1, \ldots, J, \; j \neq k$) may be tested for statistical significance. Originally proposed by Fischer and Scheiblechner (1970) in the context of examining data–model fit, the following index can be used for that purpose (see also Wright & Masters, 1982):

$$t_{j,k} = \frac{\hat{\alpha}_j - \hat{\alpha}_k}{(SE_j^2 + SE_k^2)^{1/2}}, \tag{13}$$

where SE_j and SE_k are the standard errors associated with severity measures $\hat{\alpha}_j$ and $\hat{\alpha}_k$, respectively.

The statistic shown in Equation 13 is approximately distributed as a *t* statistic with $df = n_j + n_k - 2$ (n_j and n_k are the number of ratings provided by raters *j* and *k*, respectively). For example, from Table 5 we see that the severity measures for Rater 14 and Rater 15 differed by 0.62 logits. Using Equation 13, this severity difference proved to be statistically significant; that is, $t_{14,15}(211) = 2.18$, $p < .01$.

Another group-level separation statistic is the *rater separation ratio*. This statistic gives the spread of the rater severity measures relative to the precision of those measures; that is, the closer its value is to 0, the more similar the raters are to each other in terms of their severity. Specifically, the rater separation ratio G_J is expressed as a ratio of the "true" standard deviation of rater severity measures (i.e., the standard deviation of rater severity measures corrected for measurement error; $SD_{t(J)}$) to the average rater measurement error (i.e., the "root mean-square error" associated with severity measures; $RMSE_J$):

$$G_J = SD_{t(J)} \big/ RMSE_J . \tag{14}$$

The "true" variance of rater severity measures (i.e., the square of the numerator in Equation 14) is the difference between the observed variance of rater severity measures and the average of the rater measurement error variances; $RMSE_J$ is the square root of these average error variances.

G_J indicates the spread of rater severity measures in measurement error units. The higher the value of this statistic, the more spread out the raters are on the severity scale. For our sample data, $G_J = 6.42$. This means that the rater severity differences were more than six times greater than the error of measurement.

Using the rater separation ratio, one can calculate the *rater separation index,* which is the number of statistically distinct levels of rater severity in a given sample of raters, separated by at least three measurement error units. The rater separation index (also called the *number of strata index*; Wright & Masters, 1982, 2002) is given by:

$$H_J = (4SD_{t(J)} + RMSE_J) \big/ (3RMSE_J) = (4G_J + 1)/3. \tag{15}$$

For example, a rater separation index of 3.0 would suggest that raters can be separated into three statistically distinct groups. By the same logic, when all raters were exercising a similar level of severity and thus could be considered as functioning interchangeably, a separation index close to 1 would be observed. The current analysis yielded a rater separation index of 8.89. That is, nearly 9 levels (classes, strata) of rater severity were distinguishable in this particular sample of raters.

The last separation statistic to be considered here is the *reliability of rater separation index*. This index provides information about how well the elements within the rater facet are separated in order to define reliably the facet. Rater separation reliability can be computed as a ratio of the "true" variance of rater severity measures (i.e., $SD^2_{t(J)}$) to the observed variance of rater severity measures (i.e., $SD^2_{o(J)}$):

$$R_J = SD^2_{t(J)} \big/ SD^2_{o(J)} = G_J^2 \big/ (1 + G_J^2). \tag{16}$$

Thus, R_J represents the proportion of the observed variance of rater severity measures that is *not* due to measurement error.

Note that, unlike interrater reliability, which (broadly speaking) is a measure of how *similar* rater severity measures are, rater separation reliability is a measure of how *different* rater severity measures are. In other words, when raters within a group exercise a highly similar degree of severity, rater separation reliability will be close to 0. By contrast, when raters within a group exercise a highly dissimilar degree of severity, rater separation reliability will be close to 1. Not surprisingly, in the present analysis, rater separation reliability was as high as .98, attesting to a marked heterogeneity of rater severity measures.

5.3.4 Rater severity and interrater reliability

We are now in a position to relate the severity measures estimated in the present MFRM analysis to the rater consensus and consistency indices computed earlier (see Table 1). Particularly instructive are the severity measures for those raters belonging to one of the three rater pairs discussed in Section 4.2.4.

Figure 2 and Table 5, respectively, have shown that Raters 13 and 16 are located at the severe end of the logit scale. In fact, these raters were the two most severe raters in the group. The high reliability indices (consensus and consistency) observed for this rater pair (see Table 1) were thus due to their similar tendencies to rate examinee performance very harshly. Clearly, then, this is an instance of high reliability that must not be interpreted as evidence of accurate ratings; that is, on average, Raters 13 and 16 strongly *underestimated* the language proficiency of examinees in their respective samples, as compared to the other raters.

Considering the location of Rater 13 in relation to that of Rater 03, it is evident that things are quite different. In fact, Rater 03 turned out to be one of the most lenient raters in the group. Hence, it is not at all surprising that these two raters disagreed in the majority of cases (see Table 3). The low consensus values reported for these raters were obviously due to pronounced severity differences. At the same time, this particular case demonstrates that the consistency indices which yielded moderately high values (see Table 1) actually worked to conceal the striking difference in both raters' views of examinee performance.

Finally, the severity measures that were estimated for Raters 01 and 14 similarly point to a strong rater severity effect. Whereas Rater 14 was among the more severe raters in the group, Rater 01 was a highly lenient one. As a result, at least part of the low reliability indices (consensus and consistency) observed for this rater pair (see Table 1) could be accounted for by marked severity differences between these two raters.

The considerable degree of rater variability and the ensuing problems for the interpretation of interrater reliability indices are by no means specific to the sample data studied here, neither are they specific to the TestDaF writing section. Rather, as discussed at the beginning of this chapter, rater variability is a general, notorious problem of rater-mediated assessments.

Reviewing the implications that pronounced differences in rater severity have for rater training, McNamara (1996) recommended

> . . . to accept that the most appropriate aim of rater training is to make raters internally consistent so as to make statistical modelling of their characteristics possible, but beyond this to accept variability in stable rater characteristics as a fact of life, which must be compensated for in some way . . . (p. 127)

Indeed, well-designed rater training can be effective in terms of increasing *within-rater consistency* (see, e.g., Elder et al., 2005; Weigle, 1998; Wigglesworth, 1993). Given that raters demonstrate sufficiently high degrees of internal consistency, an efficient way to compensate for rater differences in severity is to compute, for each examinee, a fair score based on many-facet Rasch model parameter estimates. How this can be done is discussed in the next section.

5.4 Examinee measurement results

Rater severity differences in the order revealed here can have important consequences for examinees. Particularly, when examinees' scores lie in critical decision-making regions of the score distribution, the final scores awarded to examinees may be affected by even small adjustments for differences in rater severity (see, for a detailed discussion, Myford et al., 1996).

To illustrate, consider examinees' observed or raw scores, computed as the average of ratings across the two raters involved, in relation to these examinees' adjusted or fair scores, computed on the basis of MFRM parameter estimates. In a way analogous to the computation of fair averages for raters, examinee fair scores compensate for rater severity differences. That is, for each examinee, there is an expected rating that would be obtained from a rater with an average level of severity. The reference group for computing this average severity level is the total group of raters included in the analysis.[9]

An observed average, or observed score, for examinee *n* is that examinee's mean rating across all raters and criteria involved in producing each rating:

9 If there is reason to believe that the reference group of raters as a whole has been unduly harsh or lenient, or if only a subgroup of raters with known level of severity or leniency has been available, either benchmark ratings or group anchoring procedures can be used to compensate for any group-level severity effects (see also Linacre, 2008).

$$M_{O(n)} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} x_{nij}}{I \cdot J}. \tag{17}$$

To compute a fair average, or fair score, for examinee *n*, the parameter estimates of all elements of the other facets that participated in producing the ratings, except for examinee *n*'s proficiency parameter, are set to their mean values. In the present three-facet example, Equation 1 becomes

$$\ln\left[\frac{p_{nk}}{p_{nk-1}}\right] = \theta_n - \beta_M - \alpha_M - \tau_k, \tag{18}$$

where p_{nk} is the probability of examinee *n* receiving a rating in category *k* across all raters and criteria, and β_M and α_M are the mean criterion difficulty and the mean rater severity measures, respectively.

The fair average (or expected score) for examinee *n* is then given as:

$$M_{F(n)} = \sum_{r=0}^{m} r p_{nr}. \tag{19}$$

Analogous to Equation 10, use of Equation 19 transforms the examinee proficiency measures, the range of which is theoretically infinite, back to the raw-score scale, which, in the present application, has a lower bound of 2 (rating category *below TDN 3*) and an upper bound of 5 (rating category *TDN 5*).

Equation 19 defines the so-called *test characteristic function* (TCF); the graphical representation of this function is the *test characteristic curve* (TCC; see, e.g., de Ayala, 2009; Yen & Fitzpatrick, 2006). The TCC is an S-shaped curve (i.e., an ogive), illustrating the fact that the functional relationship between the measures and the fair (expected) scores is nonlinear.

Fair averages for examinees greatly help to illustrate the deleterious consequences that may ensue when raw scores are taken at face value. Table 6 displays a portion of the measurement results for examinees selected from the present analysis.

A case in point is Examinee 111. This examinee proved to be highly proficient (5.38 logits, *SE* = 0.81), and the six ratings he or she received showed satisfactory model fit. The observed average was 4.33. Using the TestDaF rounding rule (see Section 4.2.2), the final level awarded would have been *TDN 4*. By contrast, the fair average computed on the basis of the estimated parameters of the facets model was 4.84, yielding final level *TDN 5* (i.e., the highest proficiency level on the TDN scale). Much the same *upward adjustment* of final TDN level would have occurred with Examinee 091. Conversely, Examinees 059 and 230 would have experienced a *downward adjustment*, if the respective fair averages were taken into account, as opposed to the observed averages. In the remaining six cases, no change in TDN level would have occurred if fair, instead of observed averages, were to provide the basis for level assignments.

Precisely what are the reasons for upward or downward adjustments of examinee proficiency levels? The data summarized in Table 7 help to provide the answer.

Table 7 shows which raters had been assigned to each of the 10 examinees listed in the previous table. In addition, each rater's severity measure and the specific ratings he or she provided on each of the three criteria are presented. Also included are the TDN levels computed by means of model parameters (fair averages) or by means of the simple averaging rule (observed averages).

Now it is plain to see that the upward adjustment of the TDN level for Examinee 111 came about because this examinee had happened to be rated by Raters 13 and 16, which, as we know from the analysis, were the two most severe raters in the group. Given that both raters provided consistent ratings (see Table 5), it can be concluded that these two raters strongly underestimated the writing proficiency of that examinee, as compared to the other raters. This underestimation was compensated for by using

Table 6. Measurement Results for the Examinee Facet (Illustrative Examples)

Examinee	*Proficiency Measure*	*SE*	*Infit*	*Outfit*	*Fair Average*	*Obs. Average*	*Number of Ratings*
111	5.38	0.81	0.94	0.95	4.84	4.33	6
239	4.14	0.91	1.00	0.89	4.60	4.67	6
091	4.12	0.83	1.24	1.23	4.59	4.33	6
059	2.31	0.83	0.91	0.85	4.12	4.50	6
032	2.17	0.77	0.87	0.86	4.09	3.50	6
213	1.29	0.79	0.70	0.70	3.88	3.83	6
153	0.41	0.80	0.91	0.89	3.65	3.67	6
230	–0.46	0.80	0.39	0.36	3.39	3.83	6
198	–1.78	0.78	1.16	1.16	3.02	3.17	6
149	–2.67	0.78	0.40	0.39	2.78	2.83	6

Note. *SE* = Standard error. Infit and outfit are mean-square statistics.

Table 7. Combined Measurement Results for Examinees and Raters (Illustrative Examples)

Examinee	*Rater*	*Severity Measure*	*Criterion Ratings GI, TT, LR (TDN*)*	*Fair Average (TDN)*	*Obs. Average (TDN)*
111	13 16	2.09 2.40	4, 4, 4 (4) 5, 5, 4 (5)	4.84 (5)	4.33 (4)
239	08 12	0.14 –1.00	5, 4, 5 (5) 5, 4, 5 (5)	4.60 (5)	4.67 (5)
091	14 08	1.83 0.14	4, 4, 3 (4) 5, 5, 5 (5)	4.59 (5)	4.33 (4)
059	12 03	–1.00 –2.01	5, 5, 4 (5) 5, 4, 4 (4)	4.12 (4)	4.50 (5)
032	13 16	2.09 2.40	4, 4, 4 (4) 3, 3, 3 (3)	4.09 (4)	3.50 (4)
213	13 03	2.09 –2.01	3, 3, 3 (3) 5, 5, 4 (5)	3.88 (4)	3.83 (4)
153	01 14	–2.23 1.83	4, 4, 4 (4) 3, 3, 4 (3)	3.65 (4)	3.67 (4)
230	07 10	–2.24 –1.02	4, 4, 4 (4) 4, 3, 4 (4)	3.39 (3)	3.83 (4)
198	15 07	1.21 –2.24	2, 2, 3 (2) 4, 4, 4 (4)	3.02 (3)	3.17 (3)
149	17 11	–0.57 0.16	3, 3, 3 (3) 3, 2, 3 (3)	2.78 (3)	2.83 (3)

Note. GI = Global impression. TT = Treatment of the task. LR = Linguistic realization. TDN* = Rater-provided TDN level. TDN = Final TDN level. TDN levels range from 2 (*below TDN 3*, lowest proficiency level) to 5 (*TDN 5*, highest proficiency level).

the examinee's fair average.[10] Likewise, the downward adjustment of the TDN level for Examinee 059 came about because this examinee had happened to be rated by Raters 12 and 03, which, as we again know from the analysis, were among the most lenient raters in the group. That is, these two raters overestimated the writing proficiency of that examinee, as compared to the other raters (for a related discussion of score adjustments, see Coniam, 2008).

There are six cases in which no upward or downward adjustment occurred, yet these cases are revealing about the impact of the severity effect on the ratings provided. For example, Examinee's 213 fair and observed averages were highly similar (3.88 vs. 3.83), resulting in the same final TDN level (i.e., *TDN 4*). However, the raters involved (Raters 13 and 03) were located at opposing ends of the severity dimension, with TDNs as provided by these raters differing by no less than two TDN levels. Thus, in cases like this, pronounced between-rater severity differences cancelled each other out, making the net result look pretty much like a fair TDN level. It is not hard to imagine what the result would have been if Examinee 213 had happened to be rated by Rater 13 and Rater 16 (which had been Examinee 111's bad luck).

The overall effect of adjusting scores for variations in rater severity across all examinees can be judged by creating a scatter diagram that plots examinee fair scores against their concomitant rater-dependent observed scores (Lunz, Wright & Linacre, 1990; McNamara, 1996). Figure 3 displays the *score adjustment diagram* obtained for the present sample of examinees.

Fair and observed averages were highly correlated (Pearson's $r = .96$, $p < .001$; Kendall's tau-b = .86, $p < .001$). Yet, in a notable number of cases, the differences between both kinds of averages were large enough as to have a critical impact on the assignment of final TDN levels. For example, given an observed average of 3.50, fair averages ranged from 3.04, suggesting *TDN 3*, to 4.09, suggesting *TDN 4*. Actually, in 53 cases (i.e., 17.3% of the sample) observed and fair averages would have led to differences in TDN assignments by exactly one TDN level (Cohen's weighted kappa = .81). The MFRM analysis therefore prevented a possible misclassification of about one-sixth of the examinees.

For the purposes of illustration, in Figure 3 dashed lines are drawn at fair and observed averages equal to 3.50. The intersection of the two lines creates four regions (or quadrants). The top-right and bottom-left regions show correctly classified examinees, that is, these examinees would have been assigned to levels *TDN 3* or *TDN 4* by both fair and observed average (levels *below TDN 3* and *TDN 5* are not considered in this example).

By contrast, the top-left and bottom-right regions show incorrectly classified examinees. Thus, examinees located in the top-left region would have been assigned to *TDN 4* by observed average but to *TDN 3* by fair average. This corresponds to a downward adjustment in 22 cases (note that some examinees had identical combinations of fair and observed average and thus are represented by dots printed on top of each other). Conversely, examinees located in the bottom-right region would have been assigned to *TDN 3* by observed average but to *TDN 4* by fair average. This corresponds to an upward adjustment in 3 cases. Across TDN levels, there would have been 41 downward adjustments and 12 upward adjustments.

Generally speaking, the horizontal spread of fair averages corresponding to each observed average shows the degree to which differences in rater severity obscure the meaning of an observed score, and the vertical spread of observed scores corresponding to each fair score shows the range of observed scores that an examinee of any given proficiency might receive depending on the rater or raters that happened to rate him or her. How many downward or upward adjustments result in a given analysis, and which levels are affected by each kind of adjustment, depends on (a) the distribution of the observed scores in the sample of examinees, (b) the distribution of the severity measures within the group of raters awarding the scores, and (c) the number of categories contained in the rating scale and consistently used by the raters (the higher this number, the more adjustments are likely to result).

10 With respect to the specific case of Examinee 111, Rater 16 provided less harsh ratings than Rater 13, although the severity estimate for Rater 16 was higher than that of Rater 13. Yet, both raters' overall severity estimates did not differ significantly. Moreover, Rater 16 tended to provide a greater number of harsh ratings at lower proficiency levels, particularly at level *TDN 3* (see Table 2, Section 4.2).

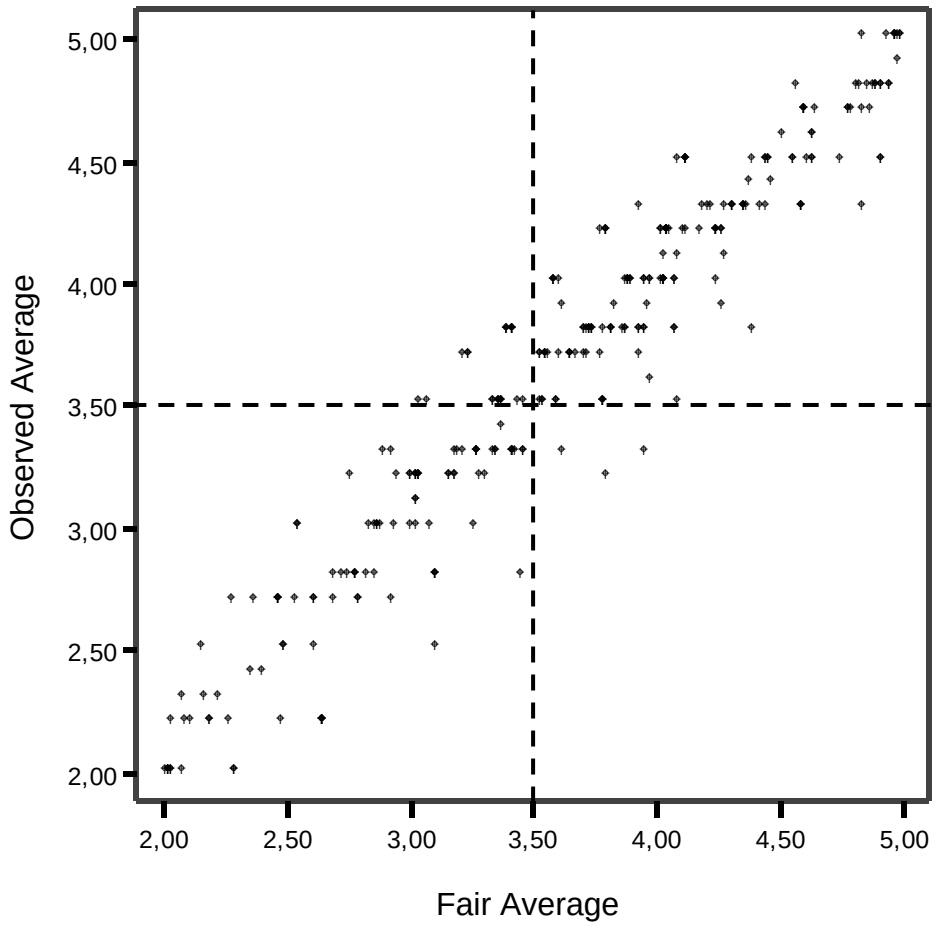

Figure 3. Score adjustment diagram. The diagram shows the relationship between fair, adjusted scores (horizontal axis) and observed, unadjusted scores (vertical axis).

5.5 Criterion measurement results

As mentioned earlier, raters utilized three criteria when scoring examinee performance: *global impression*, *treatment of the task*, and *linguistic realization*. In the present analysis, this set of criteria formed the third facet. Table 8 presents the measurement results.

Considering first the estimates of criterion difficulty, it is obvious that *linguistic realization* and *treatment of the task* were fairly difficult, as compared to *global impression*. Using the approximate *t*-statistic as shown in Equation 13, the difficulty measures of the first two criteria did not differ significantly from each other, yet the difficulty measures of each of these were significantly different from the third criterion (i.e., *global impression*).

Fit indices for all three criteria stayed well within even narrow quality control limits of 0.70 and 1.30. This finding is in line with the assumption of psychometric unidimensionality of the present set of criteria (Henning, 1992; McNamara, 1996). That is, all three criteria seemed to relate to the same dimension, as assumed by the Rasch model (see also Section 6.2 for a brief discussion of the dimensionality issue).

Table 8. Measurement Results for the Criterion Facet

Criterion	*Difficulty Measure*	*SE*	*Infit*	*Outfit*	*Fair Average*	*Obs. Average*	*Number of Ratings*
LR	0.53	0.08	0.97	0.96	3.52	3.53	648
TT	0.43	0.08	1.10	1.07	3.55	3.55	648
GI	–0.97	0.08	0.90	0.91	3.93	3.88	648

Note. LR = Linguistic realization. TT = Treatment of the task. GI = Global impression. *SE* = Standard error. Infit and outfit are mean-square statistics.

5.6 Rating scale effectiveness

Another relevant issue concerns the quality of the TDN rating scale that the raters employed to evaluate examinees' essays. To examine whether the four categories on the TDN scale (i.e., *below TDN 3, TDN 3*, *TDN 4*, *TDN 5*) functioned as intended, various statistical indicators are available (for detailed guidelines, see Linacre, 2004b; see also Bond & Fox, 2007).

An important indicator refers to the average measure by rating scale category. This indicator is computed as the average of the examinee proficiency measures that are modeled to produce the observations in a given category. The requirement is that average measures advance monotonically with categories; that is, the higher the category, the larger the average measure. When this requirement is met, it is safe to conclude that higher ratings correspond to "more" of the variable being measured. Otherwise, the meaning of the rating scale would remain unclear and, therefore, doubt would be cast on the validity of the measurement outcomes.

Another indicator of rating scale effectiveness refers to the mean-square outfit statistic computed for each rating category. This indicator compares the average examinee proficiency measures and the expected examinee proficiency measure, that is, the examinee proficiency measure the model would predict for a given rating category if the data were to fit the model. The greater the difference between the average and the expected measures, the larger the mean-square outfit statistic will be. In general, this statistic should not exceed 2.0.

Finally, the quality of a rating scale can be judged by the ordering of the category thresholds. These thresholds should advance monotonically with categories. When they do not, that is, when the thresholds are disordered, it can be concluded that the rating scale did not function properly (for an illustrative example, see Tennant, 2004). Note, however, that the requirement of ordered thresholds is not part of the mathematical structure of the rating scale or partial credit models (see Luo, 2005; Verhelst & Verstralen, 2008).

Table 9 summarizes the findings regarding these indices. As the table shows, average measures of examinee proficiency increased as the rating categories increased. Similarly, values of the outfit mean-square statistic were equal, or very close, to the expected value of 1. Finally, there was a clear progression of scale category thresholds from –3.60 logits (i.e., the threshold between categories *below TDN 3* and *TDN 3*) to 3.70 logits (i.e., the threshold between categories *TDN 4* and *TDN 5*). Taken together, these findings strongly confirm that the TDN rating scale categories were properly ordered and working as intended.

Figure 4 provides a graphical illustration of the TDN scale functionality. Specifically, the figure shows the *category probability curves* for the four-category scale that the raters used when rating examinees on the three criteria. The horizontal axis is the examinee proficiency scale; the vertical axis gives the probability of being rated in each category. There is one curve for each category. The points along the horizontal axis at which the probability curves of two adjacent rating scale categories cross denote the category thresholds.

As can be seen, there is a separate peak for each category; that is, each category is in turn the most likely category along the latent variable. Put differently, each peak appears as a distinct "hill". Similarly, the category thresholds are nicely ordered from left to right.

Table 9. Category Statistics for the TDN Rating Scale

Category	*Absolute Frequency*	*Relative Frequency*	*Average Measure*	*Outfit*	*Threshold*	*SE*
below TDN 3	209	11%	–4.13	1.0		
TDN 3	543	30%	–1.19	0.9	–3.60	0.11
TDN 4	733	40%	1.60	1.0	–0.10	0.07
TDN 5	348	19%	4.29	1.0	3.70	0.08

Note. Outfit is a mean-square fit index. Thresholds are Rasch-Andrich thresholds. *SE* = Standard error.

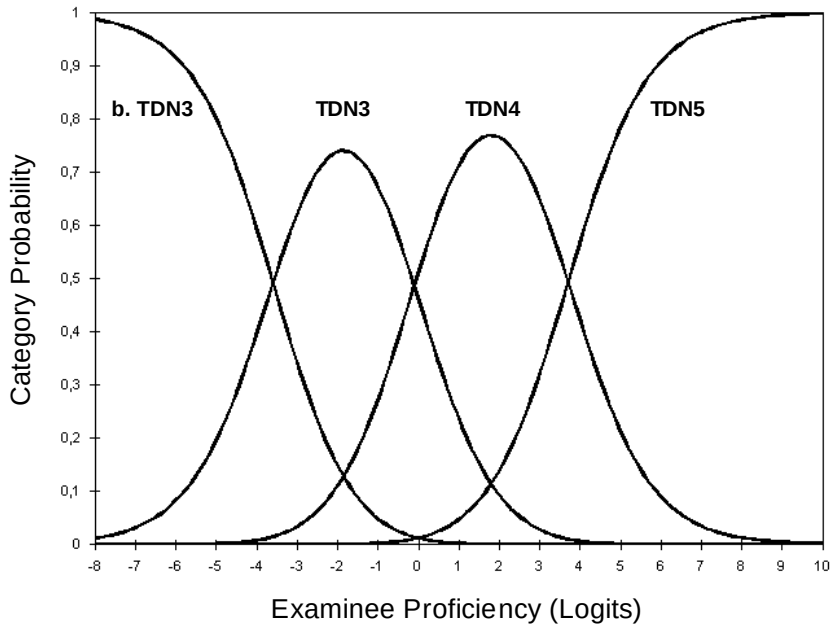

Figure 4. Category probability curves for the TDN rating scale.

5.7 Global model fit

Generally speaking, Rasch models are idealizations of empirical observations. Therefore, empirical data will never fit a given Rasch model perfectly. In other words, with a sufficiently large sample of data any model can be shown to be false (Lord & Novick, 1968). The really interesting question concerns the *practical utility* of a model; that is, we need to know whether the data fit the model usefully, and, when misfit is found, how much misfit there is and where it comes from (see also Section 5.3.1).

For the purposes of illustration, one way to assess overall data–model fit is to examine responses that are unexpected given the assumptions of the model (for a detailed discussion of various fit assessment approaches, see Fischer, 2007). According to Linacre (2008), satisfactory model fit is indicated when about 5% or less of (absolute) standardized residuals are ≥ 2, and about 1% or less of (absolute) standardized residuals are ≥ 3.

Considering the present sample, there was a total of 1,944 valid responses, that is, responses used for estimation of model parameters. Of these, 100 responses (or 5.1%) were associated with (absolute) standardized residuals ≥ 2, and 4 responses (or 0.2%) were associated with (absolute) standardized residuals ≥ 3. Overall, then, these findings would indicate satisfactory model fit.

6. MFRM Model Variations

MFRM models can be tailored to fit a variety of assessment situations, not only those involving raters. In order to provide a brief overview of model variations, I first discuss different response formats and touch upon the issue of dimensionality. Then, my focus is on polytomous responses (rating data) and different ways to model the structure of rating scales. Finally, I illustrate ways to model interactions between facets.

6.1 Response formats

The responses which are used to estimate the parameters of a particular facets model most often are polytomous, as when raters score examinee performance on rating scales. Other types of responses also suitable for a facets analysis are, of course, dichotomous responses, as when examinees take a multiple-choice vocabulary test with items scored either correct or incorrect.

Types of responses much less common in language testing contexts are binomial trials or Poisson counts (Wright & Masters, 1982). *Binomial trials* refer to situations where examinees are given a fixed number of independent attempts at an item or a task, and the number of successes or failures is counted (e.g., counts of reading or writing errors; see also Section 7.3). If the number of independent binomial trials is potentially infinite and the probability of success at each trial is small, then the resulting data may be modeled as *Poisson counts* (e.g., the number of words that an examinee can read within five minutes). Finally, even mixtures of different response types, as when dichotomous responses, polytomous responses, and binomial trials are combined in a single data set, can be used for the purposes of parameter estimation in a MFRM context.

6.2 Dimensionality

MFRM models are typically used to measure a single latent trait or dimension (e.g., examinee writing proficiency); that is, they are *unidimensional* models. Thus, when there is sufficient data–model fit, the assumption of unidimensionality is supported (see, e.g., Henning, 1992; E. V. Smith, 2002; Tennant & Pallant, 2006). In case of misfit, however, it does not follow that this assumption is to be rejected right away. Model misfit can be caused by a number of factors. Multidimensionality of the construct being measured is just one of these. Yet, when there is empirical evidence pointing to a multidimensional construct, or when theoretical considerations suggest postulating multiple latent dimensions, unidimensional models may be abandoned in favor of a multidimensional approach. For example, reading comprehension items may demand distinct cognitive operations from examinees to provide the correct answer, with each kind of operation corresponding to a different dimension, or a mathematics test may address multiple domains, including reasoning, problem solving, and spatial skills.

The basic assumption underlying the use of *multidimensional* IRT or Rasch models is that examinees vary on a number of different proficiency dimensions. In other words, an examinee's location is a point in a multidimensional space rather than a point along a single continuum. Multidimensional models are attractive because they allow for different competency profiles and learning styles, but they complicate the assignment of levels because decisions must be made (explicitly or implicitly) about the degree to which strengths along one dimension compensate for weaknesses along another dimension. In practice, the levels are defined along one dimension, and so the reported results must also lie along that one dimension. Is should also be noted that some multidimensional IRT models can produce paradoxical results; that is, getting an item correct may actually *decrease* the estimate of an examinee's proficiency in some dimension (Hooker, Finkelman & Schwartzman, in press). Surely, this calls into question the appropriateness of such models for assigning scores or levels to examinees, particularly in high-stakes testing.

Until recently, multidimensional IRT modeling approaches have only rarely been adopted in educational or language testing contexts (for some illustrative applications, see Hartig & Höhler, 2008; Liu, Wilson & Paek, 2008). General reviews were provided by Briggs and Wilson (2004), Carstensen and Rost (2007), Reckase (2007), and Rost and Walter (2006). Note that the FACETS program, which is used throughout the sample data analysis in this chapter, implements unidimensional Rasch models only.

6.3 Rating scale and partial credit models

When polytomous responses are considered, the issue of the implied structure of the rating scale comes to the fore. The facets model used in the previous sample data analysis was based on the assumption that all three criteria shared a *common* rating scale structure. That is, each category on one criterion scale (e.g., *TDN 3* on *global impression*) was assumed to be functionally equivalent to the same category on the other criterion scales (i.e., *TDN 3* on *treatment of the task* and on *linguistic realization*). Therefore, the MFRM analysis yielded only a single measurement table containing the rating scale category statistics (see Table 9), and it yielded only a single set of category probability curves (see Figure 4). The category statistics provided a summary of how the raters (as a group) used each of the four categories across the three criterion scales. Put differently, the category thresholds were assumed to be constant across the criteria.

To investigate the extent to which this assumption was actually borne out in the data, the model statement shown in Equation 1 had to be modified by changing the specification of the category coefficient term from τ_k (with a single index) to τ_{ik} (with a double index). The revised model statement would be specified as follows:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \tau_{ik}, \tag{20}$$

where all the parameters are as in Equation 1 except for the τ_{ik} term, which now represents the difficulty of scale category k relative to scale category $k - 1$ on criterion i.

Equation 20 is an expression for a *criterion-related three-facet partial credit model*. This model could also be called a *hybrid model*, since it combines a partial-credit component applied to the criteria with a rating scale component applied to the raters (Myford & Wolfe, 2003).

More specifically, the τ_{ik} term indicates that the rating scale for each criterion is modeled to have its own category structure; that is, the structure of the rating scale is allowed to vary from one criterion to another. For example, a rating of *TDN 4* on *global impression* may be more difficult (or easier) for examinees to attain relative to *TDN 5* than is a rating of *TDN 4* on *treatment of the task*. A criterion-related partial credit MFRM analysis reveals the scale structure of each individual criterion scale and thus provides information about how the group of raters used each category on each criterion.

Figure 5 shows the variable map resulting from the partial credit analysis of the sample data based on the model presented in Equation 20. The headings of the first four columns are the same as before (see Figure 2), with only slight changes in the estimates of parameters for examinees, raters, and criteria. But now each criterion has its own rating scale structure. The corresponding criterion scales are shown in columns 5 through 7 (in the order the scales were entered into the FACETS program).

Particularly interesting are the locations of the criterion-specific category thresholds (indicated by the horizontal dashed lines). Regarding *global impression* and *treatment of the task*, differences between threshold locations are negligibly small. Only the location of the threshold between categories *below TDN 3* and *TDN 3* of *linguistic realization* is placed slightly lower, as compared to the other two scales. Thus, as a group, raters used the TDN scale in much the same way across the three criteria.

Table 10 summarizes the category calibrations, or thresholds, for each criterion, as well as the means and standard deviations of the threshold estimates. This table confirms that the rating scale category calibrations were highly consistent across criteria. In each case, the differences between mean thresholds of rating scale categories were substantially larger than the corresponding standard deviations. In addition, the thresholds for each criterion were widely separated along the examinee proficiency scale. Thus, on each criterion, examinees had a high probability of being correctly classified into a rating scale category that best described their proficiency. In other words, the three criteria discriminated equally well between high and low proficiency examinees.[11]

11 On the basis of the criterion-related partial credit model, the average threshold difference computed for a particular criterion can be used as an *indirect* measure of the criterion's discrimination. Alternatively, to estimate the discrimination (or slope) parameter directly, the *generalized partial credit model* (Muraki, 1992) or the *generalized multilevel facets model* (Wang & Liu, 2007) could be employed (see also Embretson & Reise, 2000; Linacre, 2006b; Rost, 2004).

Logit	Examinee	Rater	Criterion	Rating scale for each criterion		
	High	*Severe*	*Hard*	GI	TT	LR
				(TDN 5)	(TDN 5)	(TDN 5)
8	.					
7	.					
6	*. *. *.					
5	*.					
4	** ***. **. **.			----	-----	-----
3	**. ***					
2	**. *******. ***.	16 13 14		TDN 4	TDN 4	TDN 4
1	***. ***** ***.	09 15 05				
0	****. *** ****** ****.	04 06 08 11 18 17	LR TT	----	-----	-----
-1	***. ** *****	10 12 02	GI			
-2	** **. **.	03 01 07		TDN 3	TDN 3	TDN 3
-3	**. *					
-4	***. . .			----	-----	-----
-5	* .					
-6	. .					
-7	.					
	Low	*Lenient*	*Easy*	(below 3)	(below 3)	(below 3)

Figure 5. Variable map from the many-facet partial credit analysis. Each star in the second column represents three examinees, and a dot represents one or two examinees. Scoring criteria in the fourth column are as follows: LR = linguistic realization, TT = treatment of the task, GI = global impression. The horizontal dashed lines in columns 5 through 7 indicate the category threshold measures for each of the three criterion scales.

Table 10. Rating Scale Category Calibrations for Criteria

	Global impression		*Treatment of the task*		*Linguistic realization*		*Threshold*	
Category	*Threshold*	*SE*	*Threshold*	*SE*	*Threshold*	*SE*	*M*	*SD*
TDN 3	–3.48	0.22	–3.50	0.17	–3.78	0.18	–3.59	0.17
TDN 4	–0.20	0.14	–0.20	0.13	0.05	0.13	–0.12	0.14
TDN 5	3.68	0.13	3.70	0.15	3.73	0.16	3.70	0.03

Note. Thresholds are Rasch-Andrich thresholds. *SE* = Standard error.

Two other kinds of hybrid models, which I present only briefly here, result from further varying the specification of the category coefficient. The first kind is suited to studying the way in which *each* rater used the set of criterion scales:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \tau_{jk}. \tag{21}$$

The only difference from Equation 20 is that the τ_{jk} term now represents the difficulty of scale category *k* relative to rater *j*.

Equation 21 is an expression for a *rater-related three-facet partial credit model*. This model combines a partial-credit component applied to the raters with a rating scale component applied to the criteria (Myford & Wolfe, 2003; see also Congdon & McQueen, 2000b; Wolfe, 2009). More specifically, the τ_{jk} term indicates that the rating scale for each rater is modeled to have its own category structure; that is, the structure of the rating scale is allowed to vary from one rater to another. A rater-related partial credit MFRM analysis reveals the pattern that individual raters exhibited when using the set of three criterion scales. In other words, this analysis would show how a particular rater used each category of the rating scale *across all* criteria.

Finally, to look at the way *each* rater used each category of the rating scale on *each* criterion, the partial-credit components from the models specified in Equations 20 and 21 would have to be merged:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \tau_{ijk}. \tag{22}$$

Now the τ_{ijk} term (note the triple index) represents the difficulty of scale category *k* relative to criterion *i* and rater *j*.

Equation 22 is an expression for a *criterion- and rater-related three-facet partial credit model*. This model combines a partial-credit component applied to the criteria with a partial-credit component applied to the raters (Myford & Wolfe, 2003). More specifically, the τ_{ijk} term indicates that the rating scale for each criterion *and* each rater is modeled to have its own category structure. A combined criterion- and rater-related partial credit MFRM analysis reveals the pattern that individual raters exhibited when using each of the criterion scales.

Compared to their rating scale counterparts, partial credit model variants generally require larger sample sizes in order to achieve similar stability of parameter estimates across samples. According to Linacre (1994, 2004b), usefully stable estimates may be obtained when there are at least 30 observations per element, and at least 10 observations per rating scale category. Thus, when a model such as the one specified in Equation 22 were to be used, the minimum requirement of 10 observations per category would be considerably harder to satisfy than when the rating scale model of Equation 1 were chosen.

This is because, in the partial credit model, category coefficients would have to be estimated for each criterion–rater combination separately.

6.4 Modeling facet interactions

The MFRM models discussed so far allow the researcher to single out the effect that each of a number of the facets under investigation has on the measurement results. These models do not take interactions between facets into account. That is why they are also called *main-effects models* (see, e.g., Rost & Walter, 2006; Schumacker, 1996). Yet, as mentioned in the discussion of the conceptual–psychometric framework (Section 4.3), interactions between facets may come into play and, thus, are an important issue to consider when modeling performance assessments.

Based on the parameters estimated for examinees, criteria, raters, and other facets included in a model of the kind shown in Equation 1, various interactions between facets, or differential facet functioning (DFF), can be examined. When referring to interactions involving raters, an interaction analysis is said to address *differential rater functioning* (DRF; also called *bias analysis*; see, e.g., Du et al., 1996; Engelhard, 2007a; McNamara, 1996; Myford & Wolfe, 2003; Schaefer, 2008). Depending on the number of facets considered, the analysis may address *two-way* interactions, *three-way* interactions, or even *higher-way* interactions.

With respect to the purpose of an interaction analysis, it is useful to distinguish between exploratory and confirmatory analyses. An *exploratory* interaction analysis aims at identifying systematic deviations from model expectations without any specific hypothesis in mind. That is, each and every combination of elements from two or more different facets is scanned for significant differences between observed and expected scores. The expected scores are derived from the basic MFRM model that does not include any interaction, that is, from the main-effects model. Significant differences are flagged and may then be inspected more closely. Possibly, some kind of post-hoc explanation can be reached that may in turn serve to devise a more focused interaction analysis.

Based on a theoretical rationale, on prior research, or on repeated observations, a researcher may want to test a specific interaction hypothesis; that is, a hypothesis that explicitly states which facets or which subgroups of elements of particular facets are likely to be involved in generating patterns of systematic violations of model expectations. In such a situation, a *confirmatory* interaction analysis is called for. Typically, in a confirmatory analysis some distal variable such as examinee gender or time of scoring session is considered a factor potentially exerting additional influence on the ratings.

6.4.1 Exploratory interaction analysis

To conduct an exploratory interaction analysis, the basic MFRM model is extended by adding a separate parameter that represents the interaction between the relevant facets. For example, considering again the model specified in Equation 1, the interaction between the examinee facet and the rater facet can be studied by adding an Examinee-by-Rater interaction parameter. The model statement then becomes:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \varphi_{nj} - \tau_k, \tag{23}$$

where φ_{nj} is the Examinee-by-Rater interaction parameter, also called *bias parameter* (or *bias term*).

The computer program FACETS reports a *bias statistic*, which can be used to judge the statistical significance of the size of the bias parameter estimate. That is, this statistic provides a test of the hypothesis that there is no bias apart from measurement error. The bias statistic is approximately distributed as a *t* statistic (with *df* = number of observations – 1). Referring to the bias term specified in Equation 23, the bias statistic is

$$t_{nj} = \hat{\varphi}_{nj} \big/ SE_{nj}, \tag{24}$$

where SE_{nj} is the standard error of the bias parameter estimate.

An Examinee-by-Rater interaction analysis is suited to investigate whether each rater maintained a uniform level of severity across examinees, or whether particular raters scored some examinees' performance more harshly or leniently than expected. Another two-way interaction analysis would test for patterns of unexpected ratings related to particular scoring criteria. To examine whether the combination of a particular rater and a particular criterion resulted in too harsh or too lenient scores awarded to some examinees, a three-way interaction analysis could be performed.

Drawing again on the writing performance data, Table 11 presents summary statistics showing the group-level results for these three bias analyses, that is, for the two two-way analyses (Examinee-by-Rater, Criterion-by-Rater) and for the three-way analysis (Examinee-by-Criterion-by-Rater).

Table 11 lists the total number of combinations of facet elements considered in each interaction analysis (i.e., excluding elements with extreme scores), the percentage of (absolute) *t* values equal or greater than 2, the minimum and maximum *t* values, as well as their means and standard deviations.

Whereas the percentage values for the Examinee-by-Rater and the Examinee-by-Criterion-by-Rater interactions were fairly low, a considerable number of combinations of raters and criteria were associated with substantial differences between observed and expected scores (see, for a similar finding in the TestDaF context, Eckes, 2005b). This indicates that the raters failed to keep a particular level of severity or leniency across the three criteria. They tended instead to alternate between harsher ratings on one criterion and more lenient ratings on some other criterion. Such criterion-dependent variations of rater behavior could be targeted in rater trainings.

At the individual level, that is, at the level of each individual rater, *bias control tables* and *bias control charts* (or *bias diagrams*) can be constructed for those raters exhibiting unusual rating behavior. These tables and charts also help to identify scores a rater awarded to a particular examinee that were highly unexpected, given the examinee's level of proficiency and the rater's level of severity. To illustrate, Table 12 depicts a small portion of the bias control findings for Rater 05.

For each examinee, this table lists the proficiency measure, the number of ratings (one for each criterion), the observed score (i.e., the sum of the TDNs), the expected score (based on the parameter estimates), and the average difference between observed and expected score. The last four columns are particularly relevant for an evaluation of potential rater bias related to examinees. Thus, the "Bias Measure" column gives the estimate of the interaction parameter for Rater 05 and each of the examinees (i.e., the bias in terms of the logit scale). Bias estimates greater than 0 indicate observed scores that are higher than expected based on the model (see Equation 23), while estimates less than 0 indicate observed scores that are lower than expected. Dividing the bias measure by the standard error yields the value of the bias statistic *t*. The probability associated with each *t* value is shown in the last column.

As judged by their associated probabilities, none of the bias statistic values approaches conventional levels of significance (e.g., $p \leq .05$). Yet, since statistical significance critically hinges on the number of

Table 11. Summary Statistics for the Exploratory Interaction Analysis

	Type of Interaction		
Statistic	*Examinee × Rater*	*Criterion × Rater*	*Examinee × Criterion × Rater*
N combinations	611	54	1833
% large *t*-values[a]	3.93	18.52	1.42
Minimum *t*	–3.51	–2.71	–2.72
Maximum *t*	3.30	2.78	2.56
M	–0.02	0.00	–0.05
SD	0.92	1.31	0.77

Note. [a] Percentage of absolute *t*-values ≥ 2.

Table 12. Selected Results from the Bias Analysis for Rater 05

Examinee	*Proficiency Measure*	*Number of Ratings*	*Observed Score*	*Expected Score*	*Observed – expected (Average)*	*Bias Measure*	*SE*	*t*	*p*
012	3.59	3	13	12.5	0.15	0.61	1.15	0.53	.6503
284	0.01	3	9	9.7	–0.23	–0.83	1.11	–0.75	.5318
133	–0.21	3	8	9.5	–0.50	–1.82	1.10	–1.65	.2414
295	–0.73	3	12	9.1	0.98	3.57	1.16	3.08	.0914
251	–3.79	3	6	6.8	–0.25	–0.75	1.70	–0.44	.7001

Note. *SE* = Standard error. *t* = Bias statistic.

observations (which in the present sample data is very small), it is reasonable to consider *t*'s with an absolute value of at least 2 as indicative of substantial rater bias (Engelhard, 2002; Engelhard & Myford, 2003). Following this guideline, there is one bias measure that reflects an observed score much higher than expected (marginally significant at the .10 level). This measure concerns the ratings for Examinee 295. Summed over the three criterion ratings, the observed score is almost 3 scale points higher than the expected score, resulting in an average observed–expected difference of 0.98. The bias estimate is 3.57 logits (*SE* = 1.16), which means that, from Rater's 05 perspective, Examinee 295 was 3.57 logits more proficient than his or her overall measure (i.e., –0.73 logits). Hence, this examinee's proficiency, as viewed by Rater 05, was as high as 2.84 logits.

The complete distribution of *t* values for Rater 05 plotted against the proficiency measures of the examinees scored by this rater is shown in a bias diagram (see Figure 6). For ease of interpretation, this figure also contains the upper and lower quality control limits inserted at $t = 2.0$ and $t = -2.0$, respectively. As can be seen, only two bias statistic values fall outside these limits; one of these values belongs to Examinee 295 discussed above. Since the horizontal axis refers to the examinee proficiency measures, a rough visual test can be made of whether Rater 05's bias tendency is correlated with the proficiency of the examinees he or she rated. Here, as with all the other raters studied, no such tendency was evident (Pearson's *r* for the data depicted in Figure 6 is .06, *ns*).

6.4.2 Confirmatory interaction analysis

To conduct a confirmatory interaction analysis, the basic model specification is expanded by adding at least two parameters, a new facet parameter and an interaction parameter. In this case, the first added parameter represents the facet that is the focus of the hypothesis; the second added parameter represents the interaction between that facet and some other facet already included in the model.

Hitherto, the bulk of research adopting a confirmatory approach concerned the rater facet, that is, differential rater functioning (DRF). In particular, researchers have looked at DRF related to examinee gender (see, e.g., Du & Wright, 1997; Eckes, 2005b; Engelhard & Myford, 2003) or DRF over time (also called *rater drift*; see, e.g., Congdon & McQueen, 2000a; Hoskens & Wilson, 2001; Lunz, Stahl & Wright, 1996; Wilson & Case, 2000; Wolfe, Moulder & Myford, 2001).

In a study of rater drift, *time of rating* would be considered a relevant facet. Adding a time facet to the basic model equation would allow the mean of the ratings to vary across time, but the severity of each individual rater would still be modeled as static. In order to identify individual raters who *change* their levels of severity over time, a parameter representing the interaction between the time facet and the rater facet would need to be added to the model. Changes in rating behavior that are dependent on the time of rating may manifest themselves not only in variations of rater severity, but also in variations of rater accuracy, or in variations of scale category usage (for a detailed discussion, see Wolfe, Myford, Engelhard & Manalo, 2007).

Next, I demonstrate the basic procedure of a confirmatory interaction analysis, focusing on the analysis of DRF related to examinee gender, once again using the writing performance data (see also

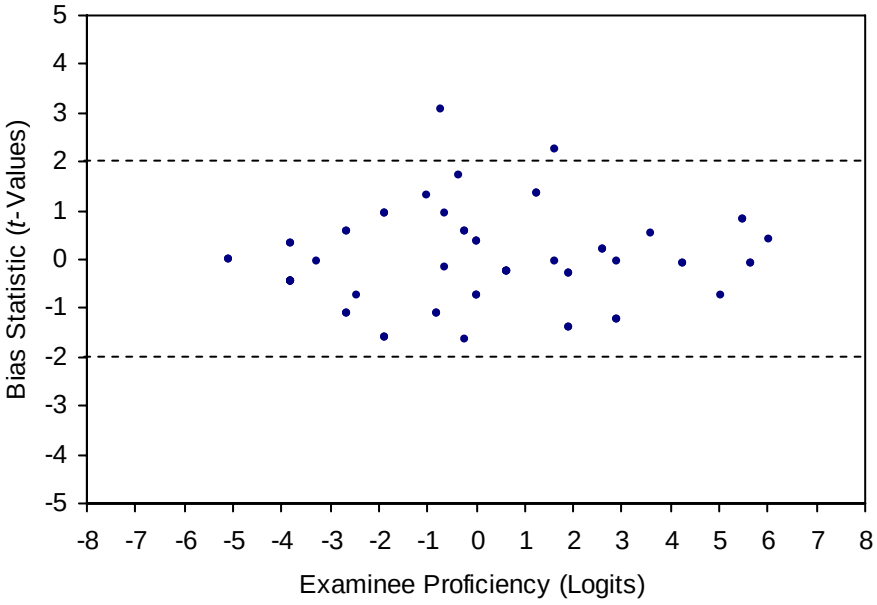

Figure 6. Bias diagram for Rater 05. Each point represents an examinee with a given proficiency measure (horizontal axis) and the associated value of the bias statistic (vertical axis). Also shown are upper and lower quality control limits (dashed lines placed at 2.0 and –2.0, respectively).

Eckes, 2005b; Engelhard & Myford, 2003).[12] The question studied was: Did any of the raters show evidence of differential severity/leniency, rating female examinees' essays (or male examinees' essays) more severely or leniently than expected, or was the ordering of raters by severity invariant across gender groups?

To answer this question, the following two terms were added to the model in Equation 1: (a) a facet term representing the examinee gender group, and (b) an interaction term representing the Rater-by-Gender Group interaction parameter. Thus, the modified model that was to provide a test of the gender bias hypothesis looked like this:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \gamma_g - \varphi_{jg} - \tau_k, \tag{25}$$

where p_{nijk} is the probability of examinee n of gender group g receiving a rating of k on criterion i from rater j, p_{nijk-1} is the probability of examinee n of gender group g receiving a rating of $k - 1$ on criterion i from rater j, γ_g is the gender facet term, and φ_{jg} is the Rater-by-Gender Group interaction term; all other terms are as in Equation 1.

In the present example, the gender bias analysis was performed by estimating proficiency measures for each group of female and male examinees along with differential proficiency measures for each and every combination of an individual rater with the respective gender group. Specifically, in order to compute the Rater-by-Gender Group interaction term, a two-step calibration was used (see Myford & Wolfe, 2003). In Step I, all parameters except φ_{jg} were estimated. In Step II, all parameters except φ_{jg}

12 Two pieces of hidden information might have contributed to the occurrence of gender bias in the sample data: (a) Each essay as well as each scoring sheet had a label attached to it, which contained, in addition to an identification number and other technical details, the examinee's full name (following the implementation of automated scanning procedures this early practice has been changed to examinee identification by number only). (b) There is some empirical evidence that raters are able to identify the gender of examinees based on handwriting (Boulet & McKinley, 2005; Emerling, 1991).

were anchored to the values estimated during the first step; then, parameter estimates and standard errors for φ_{jg} were obtained.

The null hypothesis that there is no gender bias (i.e., $\varphi_{jg} = 0$) can be tested by means of the *t* statistic introduced earlier:

$$t_{jg} = \hat{\varphi}_{jg} / SE_{jg}, \tag{26}$$

where SE_{jg} is the standard error of the gender bias parameter estimate.

A statistically significant interaction term would provide evidence for DRF. When DRF occurs, the particular Rater-by-Gender Group combination results in unexpectedly low or unexpectedly high ratings, given the rater's level of severity and the gender group's level of proficiency.

As Myford and Wolfe (2004) noted, the information provided by each of these summary statistics may be interpreted as demonstrating group-level rater differential severity/leniency only if the researcher has *prior* knowledge about whether the average measures of the gender groups should differ. Since gender differences in verbal ability have been extensively studied, though in different contexts and using different methodological approaches, at least tentative knowledge on this issue was available (see, e.g., Du & Wright 1997; Engelhard, Gordon, & Gabrielson, 1991; Hyde & Linn, 1988). For instance, in a meta-analysis covering 165 studies, Hyde and Linn (1988) found an overall mean effect size of 0.11, indicating a slight female superiority in verbal performance. More specific analyses revealed that the mean effect size was 0.09 ($p < .05$) for essay writing, and 0.33 ($p < .05$) for speech production.[13] Thus, the expectation in the present study was that females would outperform males in the writing section, albeit only to a small degree. Evidence of gender bias, therefore, would require that the calibration values for the gender facet either were very small (and not significantly different), indicating gender bias favoring males, or very large (and significantly different), indicating gender bias favoring females.

At the *group-level* analysis, the proficiency measure for females was 0.33 logits (*SE* = 0.07), that for males was –0.33 logits (*SE* = 0.07). This logit difference was statistically significant: homogeneity statistic Q_g (1) = 47.8 ($p < .01$). Therefore, it seems safe to conclude that females performed better than males, a conclusion that is in line with expectations based on prior research into the issue of gender differences in verbal ability (see, for highly similar findings in a large-scale writing assessment context, Du & Wright, 1997). Thus, there was no evidence of a group-level differential severity/leniency effect.

Deeper insight into the gender bias issue may be gained through an *individual-level* analysis. An analysis at this level indicates whether there were individual raters that displayed differential severity in their ratings. To identify such raters, a bias analysis was performed, estimating a Rater-by-Gender Group interaction term.

FACETS provided two kinds of relevant evidence, each referring to the same underlying bias/interaction information, yet from different perspectives. First, each rater was crossed with each gender group to pinpoint ratings that were highly unexpected given the pattern revealed in the overall analysis. As discussed above, any significant bias found here would provide evidence of differential rater functioning. Second, the severity of a particular rater when rating females was compared to this rater's severity when rating males. In each perspective, significant *t* values would provide evidence of individual gender bias.

Given the results of the present group-level analysis, it was not surprising that the analysis failed to find any evidence of gender bias, whichever perspective was taken. In the first (crossed) perspective, *t* values ranged from –0.81 to 0.91; in the second (pairwise) perspective *t* values ranged from –1.21 to 1.17 (all *t*'s non-significant). Nonetheless, for the purposes of illustration, Table 13 presents selected findings from the crossed individual-level analysis.

The structure of Table 13 is similar to that of Table 12, except for the added Examinee Gender column. A positive sign of the bias measure indicates that a particular rater on average awarded to a given gender group higher scores than expected on the basis of the model. Conversely, a negative sign of the bias measure indicates that a particular rater on average awarded to a given gender group lower

13 The effect size computed for each study was defined as the mean for females minus the mean for males, divided by the pooled within-gender standard deviation (see Hedges & Olkin, 1985).

Table 13. Selected Results from the Individual-Level Gender-Bias Analysis

Rater	*Severity Measure*	*Examinee Gender*	*Number of Ratings*	*Observed Score*	*Expected Score*	*Observed – expected (Average)*	*Bias Measure*	*SE*	*t*	*p*
05	1.05	Female	54	202	199.5	0.05	0.19	0.28	0.70	.4851
		Male	69	213	214.8	–0.03	–0.12	0.25	–0.46	.6502
11	0.16	Female	24	95	92.8	0.09	0.37	0.42	0.89	.3805
		Male	33	107	109.2	–0.07	–0.25	0.34	–0.75	.4570
07	–2.24	Female	111	469	472.8	–0.03	–0.17	0.21	–0.80	.4256
		Male	93	360	355.6	0.05	0.19	0.21	0.91	.3644

Note. *SE* = Standard error. *t* = Bias statistic.

scores than expected. For example, Rater 05 rated female examinees' performance slightly higher than expected and male examinees' performance slightly lower than expected; Rater 07 showed the opposite rating tendency.

Note that when *multiple* comparisons of raters are made (as in the crossed analysis presented here), critical significance levels should be adjusted to guard against falsely rejecting the null hypothesis that no biases were present (see, e.g., Engelhard, 2002). To this purpose, methods such as those based on the Bonferroni inequality (see Linacre, 2008; Myers & Well, 2003) or the Benjamini–Hochberg procedure (see Thissen, Steinberg & Kuang, 2002) can be used.

6.5 Summary of model variations

As mentioned at the beginning of this chapter, the MFRM approach does not simply refer to a single psychometric model designed for a particular purpose. Rather, MFRM is best understood as a general-purpose measurement approach that comprises a family of models each of which tailored to meet the requirements of a given assessment context. Only a few instantiations of the general approach have been discussed in the preceding sections. These and some other commonly-used models are outlined in a summary fashion in Table 14.

Model A is the rating scale model given in Equation 1 and dealt with extensively in the empirical demonstration of the MFRM approach; it is included in the table for ease of reference.

Model B represents an assessment context where raters use a single, holistic rating scale to score examinee performance on a number of different tasks (for a comparison of holistic and analytic ratings using a MFRM modeling approach, see Chi, 2001; see also Knoch, 2009).

The partial credit version of Model A is shown in the equation for Model C, with the partial credit component relating to the scoring criteria (see, for more detail, Equation 20).

Model D combines Models B and C in that criteria and tasks are included in the same equation. Moreover, this model incorporates a partial credit component that refers to both criteria and raters (but not to tasks).

Model E is typical of an investigation on examinee speaking proficiency where live interviewers present several speaking tasks, and raters score examinee performance according to a set of analytic criteria (see also the third introductory example at the beginning of the chapter).

Model F exemplifies the study of an interaction between examinees and raters (see, for more detail, Equation 23).

The final model in the summary table, Model G, includes an examinee background variable (i.e., gender) and allows the researcher to study an interaction between examinee gender and raters (see, for more detail, Equation 25).

Table 14. Examples of Measurement Models Commonly Used for the Analysis of Rater-Mediated Performance Assessments

ID	*Model*	*Facets*	*Measurement Objectives*
A	$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right]=\theta_n-\beta_i-\alpha_j-\tau_k$	Examinees, Criteria, Raters	Measurement of examinee proficiency (θ_n), criterion difficulty (β_i), and rater severity (α_j). Detailed discussion in text (see Equation 1).
B	$\ln\left[\frac{p_{nljk}}{p_{nljk-1}}\right]=\theta_n-\delta_l-\alpha_j-\tau_k$	Examinees, Tasks, Raters	Measurement of examinee proficiency (θ_n), task difficulty (δ_l), and rater severity (α_j).
C	$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right]=\theta_n-\beta_i-\alpha_j-\tau_{ik}$	Examinees, Criteria, Raters	Measurement of examinee proficiency (θ_n), criterion difficulty (β_i), and rater severity (α_j); variable structure of the rating scale for criteria. Detailed discussion in text (see Equation 20).
D	$\ln\left[\frac{p_{niljk}}{p_{niljk-1}}\right]=\theta_n-\beta_i-\delta_l-\alpha_j-\tau_{ijk}$	Examinees, Criteria, Tasks, Raters	Measurement of examinee proficiency (θ_n), criterion difficulty (β_i), task difficulty (δ_l), and rater severity (α_j); variable structure of the rating scale for criteria and raters.
E	$\ln\left[\frac{p_{nilvjk}}{p_{nilvjk-1}}\right]=\theta_n-\beta_i-\delta_l-\eta_v-\alpha_j-\tau_k$	Examinees, Criteria, Tasks, Interviewers, Raters	Measurement of examinee proficiency (θ_n), criterion difficulty (β_i), interviewer difficulty (η_v), task difficulty (δ_l), and rater severity (α_j).
F	$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right]=\theta_n-\beta_i-\alpha_j-\varphi_{nj}-\tau_k$	Examinees, Criteria, Raters	Measurement of examinee proficiency (θ_n), criterion difficulty (β_i), and rater severity (α_j); effect of the interaction between examinees and raters (φ_{nj}). Detailed discussion in text (see Equation 23).
G	$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right]=\theta_n-\beta_i-\alpha_j-\gamma_g-\varphi_{jg}-\tau_k$	Examinees, Criteria, Raters, Examinee Gender	Measurement of examinee proficiency (θ_n), criterion difficulty (β_i), and rater severity (α_j); effect of examinee gender (subgroup γ_g); effect of the interaction between raters and examinee gender (φ_{jg}). Detailed discussion in text (see Equation 25).

7. Special Issues

The MFRM approach to rater-mediated performance assessment raises a number of more specialized issues, some of which concern the design of collecting many-facet data, others relate to specific benefits that accrue from conducting a MFRM analysis. In this section, I deal with design issues first. Next, I discuss some of the benefits a MFRM approach holds for providing feedback to raters and for evaluating judgments gathered in the context of standard-setting studies. Finally, I briefly describe computer software suited to implement MFRM models.

7.1 Rating designs

In rater-mediated performance assessments, great care needs to be taken concerning the design according to which the rating data is collected. For example, when raters are to provide scores for the performance of examinees on a number of tasks (a three-facet assessment situation), questions such as the following may arise: Should all available raters score all examinees, or would it be sufficient if subsets of raters each scored a particular subset of examinees? What is a reasonable number of raters per examinee, how many examinees should each rater score, and should each rater score examinee performance on each task? With only a few raters scoring a subset of examinees, how should raters be assigned to examinees

in order to make sure that all elements of the facets involved, that is, raters, examinees, and tasks, can be represented in the same frame of reference?

To begin with, MFRM modeling is generally robust against mistakes in the implementation of a rating design. In particular, it is recommended that those in charge of the assessment program initiate the MFRM analysis as soon as data collection begins (see Linacre & Wright, 2002). This way, mistakes in the implementation of the rating design, or problematic behavior of raters, can be identified and corrected before the rating process is completed. If necessary, a conspicuous rater can be defined as "two raters", one providing ratings before remediation and the other after remediation (J. M. Linacre, personal communication, March 27, 2009).

Generally, the choice of a particular rating design depends on a mixture of measurement and practical considerations (Du & Brown, 2000; Engelhard, 1997; Hombo, Donoghue & Thayer, 2001; Myford & Wolfe, 2000; Sykes, Ito & Wang, 2008). First, other things being equal, the more data is collected, the higher the *measurement precision* of model parameters will be. For example, the larger the number of raters is per examinee, the more precise are the estimates of examinee proficiency and task difficulty.

Second, even large subsets of raters per examinee do not guard against running into serious measurement problems when the rating design does not provide for sufficient links between facet elements. This design aspect concerns the *connectedness* of the resulting data set. A connected data set is one in which a network of links exists through which every element that is involved in producing an observation is directly or indirectly connected to every other element of the same assessment context (Engelhard, 1997; Linacre & Wright, 2002; Wright & Stone, 1979). Lack of connectedness among elements of a particular facet (e.g., among raters) would make it impossible to calibrate all elements of that facet on the same scale; that is, the measures constructed for these elements (e.g., rater severity measures) could not be directly compared.

Third, in many assessment situations, particularly in large-scale assessments, considerations of *practicality* heavily narrow the choice of a rating design. Such considerations typically refer to time constraints, reasonable rater workload, and budget issues.

Table 15 illustrates the basic structure of rating designs that are suited to highlight some of the measurement and practical considerations just mentioned. These designs refer to a hypothetical assessment situation involving 10 examinees, 4 raters, and 2 tasks. Needless to say, operational rating sessions for calibrating examinees, raters, and tasks would comprise much larger sets of examinees and, possibly, raters and/or tasks, as well (see, for a detailed discussion of data collection designs in measurement contexts involving two facets, Kolen, 2007; Kolen & Brennan, 2004; Wolfe, 2000).

The first design, Rating Design A, according to which all raters score all examinees on all tasks, is an example of a *complete* or *fully crossed* design. Note that each tick mark (✓) in the design notation refers to an observation or score available for parameter estimation. A complete design is the optimum design from a measurement point of view since it leads to the highest precision of model parameter estimates possible, and to a data set that has not a single missing link. Yet, conceivably, this design is rarely, if ever, practical in real assessment situations.

More practical is Rating Design B. This is an example of an *incomplete* design, which simultaneously satisfies the measurement constraint of yielding a connected data set. Therefore, Design B is also called a *connected* (or *linked*) design: Each rater scores only a subset of examinees, and each examinee is scored by only three out of the four raters, yet all elements of all three facets are linked to each other in a common network. For example, Rater 1 is linked to Rater 3 through common ratings of Examinees 2, 4, 6, 8, and 10. Conversely, each examinee is linked to each other examinee (e.g., Examinee 1 to Examinee 2, or Examinee 3 to Examinee 10) through ratings by at least two common raters.

A further reduction in each rater's workload is achieved through incomplete Rating Design C. Each rater has to score only three or four examinees. Moreover, each examinee, except for Examinee 10, is scored by exactly one rater. Compared to complete Design A, the number of observations in Design C is reduced by 34%. Yet, since Examinee 10 is scored by all four raters, the connectedness condition is preserved.

Table 15. Illustration of Rating Designs for Three-Facet Rater-Mediated Performance Assessments

		Examinee									
Rater	*Task*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A. Complete (Fully Crossed) Design											
1	1, 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	1, 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	1, 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	1, 2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
B. Incomplete Design – Connected											
1	1, 2	✓	✓		✓	✓	✓		✓	✓	✓
2	1, 2	✓	✓	✓		✓	✓	✓		✓	✓
3	1, 2		✓	✓	✓		✓	✓	✓		✓
4	1, 2	✓		✓	✓	✓		✓	✓	✓	
C. Incomplete Design – Connected											
1	1, 2	✓			✓			✓			✓
2	1, 2					✓			✓		✓
3	1, 2			✓			✓				✓
4	1, 2		✓							✓	✓
D. Incomplete Design – Disconnected											
1	1, 2	✓	✓	✓	✓	✓	✓				
2	1, 2	✓	✓	✓	✓	✓	✓				
3	1, 2							✓	✓	✓	✓
4	1, 2							✓	✓	✓	✓
E. Incomplete, spiral design – Connected											
1	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note. Designs A through E refer to a simplified assessment situation where 10 examinees respond to 2 tasks each and 4 raters provide scores on a single, holistic rating scale. Each ✓ designates a score awarded by a given rater to an examinee's response to a task.

In contrast to Designs A through C, use of Rating Design D would result in a data set with insufficient links. This is referred to as a *disconnected* design. Though more observations are available than in Design C, the specific way of assigning raters to examinees that follows from Design D generates two *disjoint* or *disconnected subsets* of raters: Subset 1 contains Raters 1 and 2, and Subset 2 contains Raters 3 and 4. In a case like this, only measures in the same subset are directly comparable; that is, it would be misleading to compare the severity measures for Raters 1 and 3, or those for Raters 2 and 4. For example, when Rater 1 turned out to be more severe than Rater 3, it would remain unclear whether the lower scores awarded by Rater 1 were due to his or her high severity or to a low average proficiency of the examinees rated. Put differently, rater severity and examinee proficiency would be confounded.

The last design shown in Table 15 is a special variant of an incomplete, connected design that reduces the rater workload by assigning raters to score examinee performance on a subset of tasks only. Rating Design E exemplifies a *spiral* design (Hombo et al., 2001). Since performance on each task is scored by different sets of raters (Task 1 is scored by Raters 1 and 3, Task 2 is scored by Raters 2 and 4), this design is also called a *nested* design; that is, raters are nested within tasks.

As an instructive example of a real rating design consider again the sample data from the writing performance assessment described in Section 4.2.2. One of the 18 raters (Rater 06) was deliberately

chosen to rate two essays each randomly drawn from the subsets of essays already rated by each of the other 17 raters. These third ratings yielded a rating design similar in structure to Design C (see Table 15); that is, a design that is incomplete yet connected. The connecting element in Design C is Examinee 10. In the operational rating design on which the sample data analysis was based, Rater 06 provided the required connection. To illustrate, dropping this rater from the rater panel would lead to a disconnected design, with the group of raters split into three disjoint subsets as follows: Raters 01, 03, 08, 12, 13, 14, and 16 formed Subset 1; Raters 05, 07, 09, 10, 11, 15, 17, and 18 formed Subset 2; and Raters 02 and 04 formed Subset 3. Refer to Table 1 to see that raters in the same subset are linked to each other, but raters from different subsets are not. For example, there is no link between Rater 01 (Subset 1) and Rater 05 (Subset 2), whereas Rater 01 (Subset 1) is directly linked to Rater 14, and indirectly linked to Rater 03 (same subset) via Raters 14, 08, and 12.

7.2 Rater feedback

A MFRM analysis does not only provide the basis for reporting assessment results to examinees that are corrected for differences in rater severity, but also has an important role to play in rater monitoring and rater training activities. As mentioned before, research has shown that rater training can be effective in terms of increasing within-rater consistency, but that even rigorous training is unlikely to reduce differences in rater severity to any acceptable degree.

The benefits of rater training, in particular increases in within-rater consistency, may be particularly likely to be achieved through *individualized* feedback, where each rater would be given detailed information on his or her rating behavior. Results of a MFRM analysis provide a suitable basis for compiling this kind of feedback (Knoch et al., 2007; O'Sullivan & Rignall, 2007; Stahl & Lunz, 1996; Wigglesworth, 1993).

Feedback that is communicated to a particular rater may consist of one or more of the following components: (a) a severity map, showing the distribution of severity measures within the respective group of raters (particularly useful are bar graphs, where each targeted rater is clearly identified and represented by a different bar; see, e.g., Corrigan, 2007), (b) the rater's severity or leniency measure (possibly transformed to a familiar scale; see below), (c) the degree of within-rater consistency, as measured by rater infit and/or outfit indices, (d) frequency of usage of rating scale categories, and (d) quality control charts (or bias diagrams) portraying the deviations of the rater's ratings from model expectations with respect to examinees, criteria, tasks, items, or whatever other facets are considered important in the feedback process.

The rationale behind individualized feedback is that raters are construed as *independent experts* bringing individual standards and expectations to the assessment context, yet at the same willing to learn more about their rating patterns. Each rater is allowed his or her own level of severity, as long as this level is applied consistently when rating examinee performance. To this end, it is important that each piece of information conveyed to raters comes in a form that is sufficiently differentiated, easy to grasp, and supportive of each rater's efforts at becoming a proficient rater. That is, rater feedback should be encouraging and motivating, providing information where to take corrective action when necessary.

Rasch severity measures reported in logits contain decimals and negatives and can range from $-\infty$ to $+\infty$. In particular, severity measures reported in logits will be negative for one half of the rater group and positive for the other half (when, as usual, the rater facet is centered, i.e., the mean severity measure is constrained to be zero). These properties of the logit scale may be confusing to those unfamiliar with measurement results being reported in the standard unit of measurement (i.e., logits).

Therefore, when conveying severity information to raters, results may rather be reported using some sort of ordered category system (e.g., *highly lenient*, *lenient*, *average*, *severe*, *highly severe*). Consistency information could be coded in an analogous fashion (e.g., from *highly consistent* to *highly inconsistent*). More detailed and informative feedback to raters would make use of raters' fair averages. As discussed earlier (see Section 5.3.2), fair averages highlight differences between raters in direct reference to the rating scale, or scales, used during the assessment.

Alternatively, the logit scale may be linearly transformed. For example, if a scale of severity measures with the familiar range of 0 to 100 is desired, with the lowest severity measure equal to 0 and the highest severity measure equal to 100, then the following transformation of logits taken from the

rater measurement results would yield the new scale (see, e.g., E. V. Smith, 2004; Wright & Stone, 1979):

$$\hat{\alpha}_j^* = m + s \cdot \hat{\alpha}_j, \tag{27}$$

where

$$m = 0 - (s \cdot \hat{\alpha}_{j(\min)}), \tag{28}$$

and

$$s = 100 / (\hat{\alpha}_{j(\max)} - \hat{\alpha}_{j(\min)}). \tag{29}$$

In Equation 27, $\hat{\alpha}_j^*$ is the severity measure for rater *j* on the new, transformed scale, *m* is the location factor for determining the new scale origin, *s* is the spacing factor for determining the new scale unit, and $\hat{\alpha}_j$ is the severity measure for rater *j* on the old scale (i.e., the logit scale).

Finally, based on the spacing factor *s*, the standard error for the rescaled rater severity measure, that is, SE_j^*, is computed as follows:

$$SE_j^* = s \cdot SE_j. \tag{30}$$

Referring to the rater measurement results reported in Table 5, and using Equations 27 to 29, the severity measure for Rater 13 (2.09 logits, *SE* = 0.20) would become 93.31 points on the new scale (i.e., *m* = 48.27, *s* = 21.55). Rounding down would yield a severity measure of 93 points on the 0–100 scale. A value such as this one is generally much easier to communicate than the original logit value. Using Equation 30, the standard error of the rescaled severity measure for Rater 13 becomes 4.31.[14]

7.3 MFRM and standard setting

Standard setting refers to the process of establishing one or more cut scores on a test (Cizek, 2006; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Kaftandjieva, 2004; Zieky, Perie & Livingston, 2008). Cut scores are used to divide a distribution of test scores into two or more categories of performance, representing distinct levels of knowledge, competence or proficiency in a given domain. Thus, examinees may be categorized as *pass* or *fail*, or may be placed into a greater number of ordered performance categories, with labels such as *basic*, *proficient*, and *advanced*. When setting cut scores on language tests, the categories are typically taken from the CEFR six-level global scale (Council of Europe, 2001), ranging from *basic user* (A1, A2) through *intermediate user* (B1, B2) to *proficient user* (C1, C2).

According to Hambleton and Pitoniak (2006), standard setting is a "blend of judgment, psychometrics, and practicality" (p. 435). The authors characterized judgments provided by panelists as the "cornerstone" on which the resulting cut scores are based.[15] Due to the high stakes involved in many decisions that derive from the application of cut scores, it is imperative to evaluate the standard-setting process and the appropriateness of the final outcomes. Hambleton and Pitoniak suggested procedural, internal, and external sources of evidence or criteria that can be used to examine the validity of cut scores. Procedural criteria focus on issues of practicality, implementation, and documentation, internal criteria refer to consistency within and between judges, and external criteria mainly concern the replicability and the reasonableness of the cut scores.

The judge consistency issue demands particular attention. For example, judges have been shown to employ different standards when judging the difficulty of items or placing examinees into performance categories (see, e.g., Longford, 1996; Van Nijlen & Janssen, 2008). These interjudge differences need to be taken into account before determining cut scores. MFRM models are well-suited to do this. Moreover, MFRM models can be used to provide estimates of cut scores in a variety of testing and assessment

14 In an analogous fashion, examinee proficiency measures, item difficulty measures, etc., expressed in logits, can be linearly transformed to yield more familiar scales, such as the 0–100 scale.

15 In the context of standard-setting procedures raters are also called *panelists*, *judges*, or *subject matter experts* (SMEs).

contexts. In the following, I will briefly elaborate on MFRM applications to two frequently-used standard-setting procedures, the Angoff method (i.e., the unmodified variant; see Cizek & Bunch, 2007) and the bookmark method (Mitzel, Lewis, Patz & Green, 2001).

In the *Angoff method* (Angoff, 1971), judges are presented with a number of dichotomous items and asked the following question for each item: Out of 100 minimally competent examinees, how many would answer this item correctly? Viewed from a measurement perspective, the ratings obtained can be modeled as outcomes of binomial trials; that is, the number of independent trials (m) is fixed at "100", and the judges are asked to count the number of "successes", which corresponds to the number of minimally competent examinees who would answer the item correctly. The following MFRM model can be used to analyze these data:

$$\ln\left[\frac{p_{jix}}{p_{jix-1}}\right] = \alpha_j - \beta_i - \tau_x, \tag{31}$$

where

p_{jix} = probability of judge j giving a count of x on item i,

p_{jix-1} = probability of judge j giving a count of $x - 1$ on item i,

α_j = judged minimal competence for judge j,

β_i = judged difficulty for item i,

τ_x = judged difficulty of giving a count of x relative to a count of $x - 1$.

Note that parameter α_j in Equation 31 is different in meaning from the rater severity parameter discussed earlier. Now this parameter represents the severity of a particular judge's *view* of minimal competence required to answer item i correctly; that is, a severe judge would give a small count of minimally competent examinees answering that item correctly, as compared to a lenient judge. Note also that β_i in Equation 31 refers to the *judged* difficulty of item i; that is, a difficult item would be given small counts by the judges, as compared to an easy item.[16]

Based on the model given in Equation 31, statistical indicators described in previous sections of this chapter, such as separation, fit, and bias statistics, can be used to analyze the psychometric quality of the judges' ratings (see Engelhard & Anderson, 1998; Engelhard & Cramer, 1997). The variable map would provide a particularly instructive portrayal of the measurement results for the standard-setting judges and the items. As to the judge facet, the map would show the location of the minimally competent examinee as viewed by the judges, where a higher location represents a higher minimal competence required to answer items correctly (i.e., corresponding to a severe judge's view of minimal competence). With regard to the item facet, the map would show the location of each item in terms of its judged difficulty, where a higher location represents a lower count of examinees answering the item correctly.

In addition, relating the judged item difficulties to empirical item difficulties (e.g., item difficulties derived from operational test administrations), may yield evidence of the validity of the standard-setting procedure, and may thus inform the process of setting cut scores (see Baghaei, 2007; Taube, 1997; Verheggen, Muijtjens, van Os & Schuwirth, 2008).

A different kind of MFRM model is called for when evaluating a standard setting where panelists provide judgments on the *level of performance* needed to succeed on each of a number of items. The test may contain a mixture of selected-response (e.g., multiple-choice) and constructed-response items, and the judges may be asked to consider a single level or multiple levels of performance.

For example, consider a test containing 60 multiple-choice items designed to assess four performance levels (e.g., CEFR levels A2, B1, B2, and C1). In the *bookmark method* (Mitzel et al., 2001), judges would be presented with a booklet consisting of the set of 60 items, one item per page, with items ordered from easiest to hardest. Judges would be asked, for each level of performance, to place a

16 The binomial trials model reduces to the Rasch model for dichotomous data if $m = 1$; that is, $x = 0$ or 1 (Wright & Masters, 1982).

bookmark on the first page in the booklet at which they believe the probability of answering the item correctly drops below a 2/3 chance (or below a .67 probability). Thus, panelists would have to place three bookmarks in their booklet, each one identifying a cut-off between two adjacent performance levels. Judges are usually asked to repeat this marking procedure two times, each marking session constituting a separate round (see, for a detailed description of the bookmark method, Cizek & Bunch, 2007).

Each bookmark placement sorts the set of items into one of four performance categories. Bookmark placements can thus be construed as judgments or ratings of items on a four-category performance scale. A MFRM model suited to the analysis of such bookmark ratings would be:

$$\ln\left[\frac{p_{jirk}}{p_{jirk-1}}\right] = \alpha_j - \beta_i - \rho_r - \tau_k, \tag{32}$$

where

p_{jirk} = probability of judge j giving a bookmark rating of k on item i for round r,

p_{jirk-1} = probability of judge j giving a bookmark rating of $k-1$ on item i for round r,

α_j = judged performance level for judge j,

β_i = judged difficulty of item i,

ρ_r judged performance level for round r,

τ_k = judged performance standard for bookmark rating category k relative to category $k-1$.

The model given in Equation 32 is an example of a three-facet rating scale model (see Engelhard, 2007b, 2008b). Model parameters can be interpreted in much the same way as with the binomial trials model in Equation 30. Note that inclusion of the *round facet* makes it possible to study changes in between-rater differences in judged performance level as well as changes in within-rater judgment consistency from one round to the next. Finally, in the present framework, the category coefficients, τ_k, define the cut scores (possibly after removal of misfitting judges). In order to provide evidence concerning the validity of the standard-setting procedure, the measurement-based cut scores may be compared to the cut scores determined by the usual bookmark procedure.

Further examples of MFRM model applications to a range of standard-setting procedures can be found in Engelhard and Gordon (2000), Engelhard and Stone (1998), Kecker and Eckes (in press), Kozaki (2004), Lumley, Lynch and McNamara (1994), Lunz (2000), Stone (2006), and Stone, Beltyukova, and Fox (2008). In each of these studies, the MFRM modeling approach proved to be a valuable instrument for the purposes of evaluating standard-setting data and/or setting cut scores on examinations.

On a more cautionary note, an important aim of many standard-setting procedures is to reach consensus among judges before deciding on the cut scores. Yet, implicitly or explicitly forcing judges into agreement is bound to create some degree of dependence among judges that may pose problems for interpreting results from a MFRM analysis. As mentioned before, the MFRM approach basically construes raters or judges as individual experts, with each judgment providing an independent piece of information about the location of an item (or an examinee) on the latent continuum (see Linacre, 1997, 1998, 2002b). It may thus be reasonable not to perform MFRM analyses in the later stages of standard setting where judges can be assumed to gravitate toward the group mean.

More generally speaking, rater or judge dependence typically leads to a higher proportion of overfit, lowered standard errors, and a markedly widened range of parameter estimates, which makes the judgments appear more reliable than they may actually be. Linacre (2008) proposed to assess the degree of rater dependence in a given data set by means of an index that is akin to the rationale of Cohen's kappa, the so-called *Rasch-kappa index*. Other researchers have developed psychometric approaches that are suited to explicitly model this kind of dependence, such as the *rater bundle model* (Wilson & Hoskens, 2001), the *hierarchical rater model* (Patz, Junker, Johnson & Mariano, 2002; for a critique of this model, see Linacre, 2003a), and the *IRT model for multiple raters* (Verhelst & Verstralen, 2001).

The relative merits, prospects, and limitations of these and related approaches to dealing with the rater dependence issue in applied settings will need to be studied more closely in future research (see, e.g., Barr & Raju, 2003; Mariano & Junker, 2007).

7.4 MFRM software

In this chapter, MFRM analyses were conducted by means of the computer program FACETS (version 3.64; Linacre, 2008). Over the years, FACETS has gained great popularity among Rasch practitioners working in a wide range of research fields (see, e.g., the references cited in Linacre, 2008). FACETS is a highly versatile program that provides users with lots of MFRM model instantiations, analytical tools, and statistical indicators, and it offers flexible input and output functions, reporting measurement results in user-specified tables and graphical displays. At the website www.winsteps.com, interested readers will find detailed information on the program, including a free FACETS manual that contains many helpful explanations, as well as a free student/evaluation version called MINIFAC.

There are a number of other computer programs that can be used to conduct MFRM analyses, including ConQuest (Wu, Adams, Wilson & Haldane, 2007), RUMM (Andrich, Sheridan & Luo, 2004), PARSCALE (du Toit, 2003), LPCM-WIN (Fischer & Ponocny-Seliger, 2003), and the open-source software eRm (Mair & Hatzinger, 2007a, 2007b). Among other things, these programs differ in the techniques they employ for estimating model parameters. The estimation techniques implemented are: joint maximum likelihood (FACETS), marginal maximum likelihood (ConQuest, PARSCALE), pairwise conditional (RUMM), and conditional maximum likelihood (LPCM-WIN, eRm). There has been a theoretical debate about the relative merits of each technique (see, for detailed discussions, Baker & Kim, 2004; Linacre, 2004a, 2004c; Molenaar, 1995; Verhelst, 2004). In particular, joint maximum likelihood estimation has been criticized for its lack of consistency (see, e.g., Cohen, Chan, Jiang & Seburn, 2008; Molenaar, 1995). However, some authors suggested that the differences in the estimates produced by each of these techniques can be considered negligibly small for many practical purposes (see Baker & Kim, 2004; Linacre, 2004a; see also Kline, Schmidt & Bowles, 2006).

Concerning the generality of the underlying measurement approach, ConQuest stands out by using one highly general model to fit a wide variety of Rasch models: the *multidimensional random coefficients multinomial logit model* (MRCMLM; Adams, Wilson & Wang, 1997; see also Adams & Wu, 2007). An appealing feature of ConQuest is the option to perform *hierarchical model testing*. Choosing this option allows the researcher to systematically compare competing models that may each be considered appropriate for the data given. For example, let Model A specify examinees, criteria, raters, and an examinee-by-rater interaction. This model may be compared to a more parsimonious Model B (i.e., a submodel of A), created by removing the interaction term from Model A. Significantly better fit of Model A would indicate that the interaction between examinees and raters is a source of variation in the ratings that is not to be ignored.

Another important feature of ConQuest refers to the option to implement *multidimensional* Rasch models. As mentioned earlier (see Section 6.2), such models may be called for when it is reasonable to assume that there are two or more latent proficiency dimensions simultaneously addressed by a testing procedure. Furthermore, ConQuest's multidimensional option allows researchers to model local item dependence (LID). For example, LID may occur among a set of criteria on which raters provide ratings of examinee performance (Wang & Wilson, 2005a, 2005b). Multidimensional Rasch models can also be implemented through the program MULTIRA (Carstensen & Rost, 2001; see also Rost & Carstensen, 2002).

Finally, Muckle and Karabatsos (2009) have shown that the many-facet Rasch model can be considered a special case of the two-level hierarchical generalized linear model (HGLM). More generally, adopting a *multilevel* Rasch measurement perspective makes possible a number of extensions regarding the analysis of many-facet data, for example, modeling nested data structures or modeling covariates of examinee proficiency, task difficulty, and rater severity (for overviews, see Jiao, Wang & Kamata, 2007; Kamata & Cheong, 2007). HGLMs can be implemented using software such as HLM 6 (Raudenbush, Bryk, Cheong, Congdon & du Toit, 2004) or the open-source lme4 package (Doran, Bates, Bliese & Dowling, 2007; for a comparison of different HGLM software packages, see Roberts & Herrington, 2007).

Conclusion

"There is nothing more practical than a good theory" (Lewin, 1952, p. 169). Placing this famous statement in the present context, the ensuing demands are twofold: Psychometricians should develop theories and models that can be used to understand, conceptualize, and efficiently solve practical problems, such as those caused by the notoriously lacking rater agreement in rater-mediated assessment situations. Conversely, practitioners and researchers in the field of language testing and assessment should make use of available psychometric theory to provide examinees with assessment results that are as objective, valid, and fair as possible. Beyond any doubt, many-facet Rasch measurement has the potential to integrate theorists' and practitioners' interests in measuring language proficiency, and thus to meet both demands simultaneously.

Acknowledgements

I would like to acknowledge the helpful comments and suggestions from the following members of the Council of Europe's (2009) Manual Authoring Group on earlier versions of this chapter: Brian North, Sauli Takala, who also served as editor of the Reference Supplement, and Norman D. Verhelst. In addition, I received valuable feedback on the manuscript from Rüdiger Grotjahn, Klaus D. Kubinger, J. Michael Linacre, and Carol M. Myford. I would also like to express my gratitude to my colleagues at the TestDaF Institute, Hagen, Germany, for many stimulating discussions concerning the design, analysis, and evaluation of writing and speaking performance assessments.

References

Adams, R. J., Wilson, M., & Wang, W-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57–75). New York: Springer.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Andrich, D. (1998). Thresholds, steps and rating scale conceptualization. *Rasch Measurement Transactions*, *12*, 648–649.

Andrich, D., Sheridan, B. E., & Luo, G. (2004). *RUMM2020: Rasch unidimensional measurement models* [Computer software]. Perth, Western Australia: RUMM Laboratory.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*, 1–42.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

Baghaei, P. (2007). Applying the Rasch rating-scale model to set multiple cut-offs. *Rasch Measurement Transactions*, *20*, 1075–1076.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Dekker.

Barr, M. A., & Raju, N. S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods*, *6*, 15–43.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*, 49–58.

Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). Human scoring. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49–81). Mahwah, NJ: Erlbaum.

Berry, V. (2007). *Personality differences and oral test performance.* Frankfurt, Germany: Lang.

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

Boulet, J. R., & McKinley, D. W. (2005). Investigating gender-related construct-irrelevant components of scores on the written assessment exercise of a high-stakes certification assessment. *Advances in Health Sciences Education*, *10*, 53–63.

Bramley, T. (2007). Quantifying marker agreement: Terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, *4*, 22–28.

Breton, G., Lepage, S., & North, B. (2008). *Cross-language benchmarking seminar to calibrate examples of spoken production in English, French, German, Italian and Spanish with regard to the six levels of the Common European Framework of Reference for Languages (CEFR)*. Strasbourg, France: Council of Europe/Language Policy Division.

Briggs, D. C., & Wilson, M. (2004). An introduction to multidimensional measurement using Rasch models. In E. V. Smith & R.

M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 322–341). Maple Grove, MN: JAM Press.
Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt, Germany: Lang.
Carstensen, C. H., & Rost, J. (2001). *MULTIRA: A program system for multidimensional Rasch models* [Computer software]. Kiel, Germany: IPN—Leibniz Institute for Science Education.
Carstensen, C. H., & Rost, J. (2007). Multidimensional three-mode Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 157–175). New York: Springer.
Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. *Journal of Applied Measurement*, *2*, 379–388.
Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225–258). Mahwah, NJ: Erlbaum.
Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*, 213–220.
Cohen, J., Chan, T., Jiang, T., & Seburn, M. (2008). Consistent estimation of Rasch item parameters and their standard errors under complex sample designs. *Applied Psychological Measurement*, *32*, 289–310.
Congdon, P. J., & McQueen, J. (2000a). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *37*, 163–178.
Congdon, P. J., & McQueen, J. (2000b). Unmodeled rater discrimination error. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 165–180). Stamford, CT: Ablex.
Coniam, D. (2008). Problems affecting the use of raw scores: A comparison of raw scores and FACETS' fair average scores. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference May 2005* (pp. 179–190). Cambridge, UK: Cambridge University Press.
Corrigan, M. (2007). *Seminar to calibrate examples of spoken performance: Report on the analysis of rating data.* Strasbourg, France: Council of Europe/Language Policy Division.
Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press.
Council of Europe. (2009). *Relating language examinations to the common European framework of reference for languages: Learning, teaching, assessment (CEFR): A manual.* Strasbourg, France: Council of Europe/Language Policy Division.
de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: Guilford.
Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*(2).
Du, Y., & Brown, W. L. (2000). Raters and single prompt-to-prompt equating using the facets model in a writing performance assessment. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 97–111). Stamford, CT: Ablex.
Du, Y., & Wright, B. D. (1997). Effects of student characteristics in a large-scale direct writing assessment. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 1–24). Stamford, CT: Ablex.
Du, Y., Wright, B. D., & Brown, W. L. (1996, April). *Differential facet functioning detection in direct writing assessment*. Paper presented at the Annual Conference of the American Educational Research Association, New York, NY.
du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT* [Software manual]. Lincolnwood, IL: Scientific Software International.
Eckes, T. (2004). Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im „Test Deutsch als Fremdsprache" (TestDaF) [Rater agreement and rater severity: A many-facet Rasch analysis of performance assessments in the „Test of German as a Foreign Language (TestDaF)"]. *Diagnostica*, *50*, 65–77.
Eckes, T. (2005a). Evaluation von Beurteilungen: Psychometrische Qualitätssicherung mit dem Multifacetten-Rasch-Modell [Evaluation of ratings: Psychometric quality assurance via many-facet Rasch measurement]. *Zeitschrift für Psychologie*, *213*, 77–96.
Eckes, T. (2005b). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*, 197–221.
Eckes, T. (2008a). Assuring the quality of TestDaF examinations: A psychometric modeling approach. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference May 2005* (pp. 157–178). Cambridge, UK: Cambridge University Press.
Eckes, T. (2008b). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155–185.
Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt, Germany: Lang.
Elbow, P., & Yancey, K. B. (1994). On the nature of holistic scoring: An inquiry composed on e-mail. *Assessing Writing*, *1*, 91–107.
Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*, 37–64.
Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*, 175–196.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
Emerling, F. (1991). Identifying ethnicity and gender from anonymous essays. *Community College Review*, *19*, 29–33.
Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*, 19–33.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Erlbaum.
Engelhard, G. (2007a). Differential rater functioning. *Rasch Measurement Transactions*, *21*, 1124.
Engelhard, G. (2007b). Evaluating bookmark judgments. *Rasch Measurement Transactions*, *21*, 1097–1098.
Engelhard, G. (2008a). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 155–189.
Engelhard, G. (2008b). Standard errors for performance standards based on bookmark judgments. *Rasch Measurement Transactions*, *21*, 1132–1133.
Engelhard, G., & Anderson, D. W. (1998). A binomial trials model for examining the ratings of standard-setting judges. *Applied Measurement in Education*, *11*, 209–230.
Engelhard, G., & Cramer, S. E. (1997). Using Rasch measurement to evaluate the ratings of standard-setting judges. In M. Wilson, G. Engelhard, & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 97–112). Greenwich, CT: Ablex.
Engelhard, G., & Gordon, B. (2000). Setting and evaluating performance standards for high stakes writing assessments. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 3–14). Stamford, CT: Ablex.
Engelhard, G., Gordon, B., & Gabrielson, S. (1991). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. *Research in the Teaching of English*, *26*, 315–336.
Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model* (College Board Research Report No. 2003-1). New York: College Entrance Examination Board.
Engelhard, G., & Stone, G. E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement*, *58*, 179–196.
Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.
Fischer, G. H. (1995a). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15–38). New York: Springer.
Fischer, G. H. (1995b). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 131–155). New York: Springer.
Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Handbook of statistics, Vol. 26, pp. 515–585). Amsterdam: Elsevier.
Fischer, G. H., & Ponocny-Seliger, E. (2003). *Structural Rasch modeling: Handbook of the usage of LPCM-WIN 1.0* [Software manual]. Groningen, The Netherlands: Science Plus Group.
Fischer, G. H., & Scheiblechner, H. H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch [Algorithms and programs for Rasch's probabilistic test model]. *Psychologische Beiträge*, *12*, 23–51.
Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: Wiley.
Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S. A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75–98). New York: Longman.
Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Education.
Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.
Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, *12*, 1–9.
Harasym, P. H., Woloschuk, W., & Cunning, L. (2008). Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Advances in Health Sciences Education*, *13*, 617–632.
Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology*, *216*, 89–101.
Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77–89.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*, 1–11.
Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (2001). *A simulation study of the effect of rater designs on ability estimation* (Research Report, RR-01-05). Princeton, NJ: Educational Testing Service.
Hooker, G., Finkelman, M., & Schwartzman, A. (in press). Paradoxical results in multidimensional item response theory. *Psychometrika*.
Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement*, *38*, 121–145.
Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*, 64–86.
Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*, 403–424.
Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*, 53–69.
Jiao, H., Wang, S., & Kamata, A. (2007). Modeling local item dependence with the hierarchical generalized linear model. In E. V. Smith & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 390–404). Maple Grove, MN: JAM Press.
Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* New York: Guilford.
Kaftandjieva, F. (2004). Standard setting. In S. Takala (Ed.), *Reference supplement to the preliminary pilot version of the manual*

for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section B). Strasbourg, France: Council of Europe/Language Policy Division.
Kamata, A., & Cheong, Y. F. (2007). Multilevel Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 217–232). New York: Springer.
Kecker, G., & Eckes, T. (in press). Putting the Manual to the test: The TestDaF–CEFR linking project. In W. Martyniuk (Ed.), *Linking tests to the CEFR: Case studies and reflections on using the Council of Europe's draft Manual for relating language examinations to the CEFR.* Cambridge, UK: Cambridge University Press.
Kempf, W. F. (1972). Probabilistische Modelle experimentalpsychologischer Versuchssituationen [Probabilistic models of designs in experimental psychology]. *Psychologische Beiträge*, *14*, 16–37.
Kline, T. L., Schmidt, K. M., & Bowles, R. (2006). Using LinLog and FACETS to model item components in the LLTM. *Journal of Applied Measurement*, *7*, 74–91.
Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). New York: Springer.
Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.
Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*, 275–304.
Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26–43.
Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*, 3–31.
Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, *21*, 1–27.
Kubinger, K. D. (2005). Psychological test calibration using the Rasch model: Some critical suggestions on traditional approaches. *International Journal of Testing*, *5*, 377–394.
Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement*, *69*, 232–244.
Lamprianou, I. (2006). The stability of marker characteristics across tests of the same subject and across subjects. *Journal of Applied Measurement*, *7*, 192–205.
Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*, 72–107.
Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education/Praeger.
LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852.
Lewin, K. (1952). *Field theory in social science: Selected theoretical papers.* London: Tavistock.
Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.
Linacre, J. M. (1994). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions*, *7*, 328.
Linacre, J. M. (1996). True-score reliability or Rasch validity? *Rasch Measurement Transactions*, *9*, 455.
Linacre, J. M. (1997). Investigating judge local independence. *Rasch Measurement Transactions*, *11*, 546–547.
Linacre, J. M. (1998). Rating, judges and fairness. *Rasch Measurement Transactions*, *12*, 630–631.
Linacre, J. M. (2002a). Facets, factors, elements and levels. *Rasch Measurement Transactions*, *16*, 880.
Linacre, J. M. (2002b). Judge ratings with forced agreement. *Rasch Measurement Transactions*, *16*, 857–858.
Linacre, J. M. (2002c). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.
Linacre, J. M. (2003a). The hierarchical rater model from a Rasch perspective. *Rasch Measurement Transactions*, *17*, 928.
Linacre, J. M. (2003b). Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions*, *17*, 918.
Linacre, J. M. (2004a). Estimation methods for Rasch measures. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 25–47). Maple Grove, MN: JAM Press.
Linacre, J. M. (2004b). Optimizing rating scale category effectiveness. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258–278). Maple Grove, MN: JAM Press.
Linacre, J. M. (2004c). Rasch model estimation: Further topics. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 48–72). Maple Grove, MN: JAM Press.
Linacre, J. M. (2006a). Demarcating category intervals. *Rasch Measurement Transactions*, *19*, 1041–1043.
Linacre, J. M. (2006b). Item discrimination and Rasch-Andrich thresholds. *Rasch Measurement Transactions*, *20*, 1054.
Linacre, J. M. (2008). *Facets Rasch model computer program* [Software manual]. Chicago: Winsteps.com.
Linacre, J. M., Engelhard, G., Tatum, D. S., & Myford, C. M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research*, *21*, 569–577.
Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, *3*, 484–509.
Liu, O. L., Wilson, M., & Paek, I. (2008). A multidimensional Rasch analysis of gender differences in PISA mathematics. *Journal of Applied Measurement*, *9*, 18–35.
Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*, 493–504.
Longford, N. T. (1996). Reconciling experts' differences in setting cut scores for pass-fail decisions. *Journal of Educational and Behavioral Statistics*, *21*, 203–213.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.
Lumley, T. (2005). *Assessing second language writing: The rater's perspective.* Frankfurt, Germany: Lang.
Lumley, T., Lynch, B. K., & McNamara, T. F. (1994). A new approach to standard-setting in language assessment. *Melbourne Papers in Language Testing*, *3*, 19–39.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 54–71.
Lunz, M. E. (2000). Setting standards on performance examinations. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 181–199). Stamford, CT: Ablex.
Lunz, M. E. (2007). An example of grader consistency using the multi-facet model. *Rasch Measurement Transactions*, *21*, 1102.
Lunz, M. E., & Linacre, J. M. (1998). Measurement designs using multifacet Rasch modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 47–77). Mahwah, NJ: Erlbaum.
Lunz, M. E., Stahl, J. A., & Wright, B. D. (1996). The invariance of judge severity calibrations. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 99–112). Norwood, NJ: Ablex.
Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*, 331–345.
Luo, G. (2005). The relationship between the rating scale and partial credit models and the implication of disordered thresholds of the Rasch models for polytomous responses. *Journal of Applied Measurement*, *6*, 443–455.
Mair, P., & Hatzinger, R. (2007a). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, *49*, 26–43.
Mair, P., & Hatzinger, R. (2007b). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9).
Mariano, L. T., & Junker, B. W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. *Journal of Educational and Behavioral Statistics*, *32*, 287–314.
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.
McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.
McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education*, *6*(42).
McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.
McNamara, T. F. (2000). *Language testing*. Oxford, UK: Oxford University Press.
McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
Micko, H. C. (1969). A psychological scale for reaction time measurement. *Acta Psychologica*, *30*, 324–335.
Micko, H. C. (1970). Eine Verallgemeinerung des Meßmodells von Rasch mit einer Anwendung auf die Psychophysik der Reaktionen [A generalization of Rasch's measurement model with an application to the psychophysics of reactions]. *Psychologische Beiträge*, *12*, 4–22.
Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Erlbaum.
Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 39–51). New York: Springer.
Muckle, T. J., & Karabatsos, G. (2009). Hierarchical generalized linear models for the analysis of judge ratings. *Journal of Educational Measurement*, *46*, 198–219.
Mun, E. Y. (2005). Rater agreement – weighted kappa. In B. S. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1714–1715). New York: Wiley.
Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.
Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.
Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Erlbaum.
Myford, C. M., Marr, D. B., & Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the Test of Written English* (TOEFL Research Report No. 95-40). Princeton, NJ: Educational Testing Service.
Myford, C. M., & Wolfe, E. W. (2000). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs* (TOEFL Technical Report, TR-15). Princeton, NJ: Educational Testing Service.
Myford, C. M., & Wolfe, E. W. (2002). When raters disagree, then what? Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement*, *3*, 300–324.
Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386–422.
Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*, 189–227.
North, B. (2000). *The development of a common framework scale of language proficiency.* New York: Lang.
North, B. (2008). The CEFR levels and descriptive scales. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference May 2005* (pp. 21–66). Cambridge, UK: Cambridge University Press.
North, B., & Jones, N. (2009). *Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling.* Strasbourg, France: Council of Europe/Language Policy Division.
North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*, 217–262.
O'Neill, T. R., & Lunz, M. E. (2000). A method to study rater severity across several administrations. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 135–146). Stamford, CT: Ablex.
Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.
O'Sullivan, B. (2008). *Modelling performance in tests of spoken language.* Frankfurt, Germany: Lang.
O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS Writing Module. In L.

Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 446–478). Cambridge, UK: Cambridge University Press.
Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*, 341–384.
Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Original work published 1960)
Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM 6: Hierarchical linear and nonlinear modeling* [Software manual]. Lincolnwood, IL: Scientific Software International.
Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Handbook of statistics, Vol. 26, pp. 607–642). Amsterdam: Elsevier.
Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder et al. (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 82–96). Cambridge, UK: Cambridge University Press.
Roberts, J. K., & Herrington, R. (2007). Demonstration of software programs for estimating multilevel measurement model parameters. In E. V. Smith & R. M. Smith (Eds.), *Rasch measurement: Advanced and specialized applications* (pp. 303–328). Maple Grove, MN: JAM Press.
Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*, 271–282.
Rost, J. (2001). The growing family of Rasch models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 25–42). New York: Springer.
Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion* [Textbook test theory, test construction] (2nd ed.). Bern, Switzerland: Huber.
Rost, J., & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. *Applied Psychological Measurement*, *26*, 42–56.
Rost, J., & Langeheine, R. (1997). A guide through latent structure models for categorical data. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 13–37). Münster, Germany: Waxmann.
Rost, J., & Walter, O. (2006). Multimethod item response theory. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 249–268). Washington, DC: American Psychological Association.
Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*, 465–493.
Schumacker, R. E. (1996, April). *Many-facet Rasch model selection criteria: Examining residuals and more*. Paper presented at the Annual Conference of the American Educational Research Association, New York, NY.
Schumacker, R. E., & Smith, E. V. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement*, *67*, 394–409.
Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, *15*, 188–211.
Shoukri, M. M. (2004). *Measures of interobserver agreement*. Boca Raton, FL: Chapman & Hall/CRC.
Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*, 205–231.
Smith, E. V. (2004). Metric development and score reporting in Rasch measurement. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 342–365). Maple Grove, MN: JAM Press.
Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). Maple Grove, MN: JAM Press.
Stahl, J. A., & Lunz, M. E. (1996). Judge performance reports: Media and message. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 113–125). Norwood, NJ: Ablex.
Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Los Angeles: Sage.
Stone, G. E. (2006). Whose criterion standard is it anyway? *Journal of Applied Measurement*, *7*, 160–169.
Stone, G. E., Beltyukova, S., & Fox, C. M. (2008). Objective standard setting for judge-mediated examinations. *International Journal of Testing*, *8*, 180–196.
Su, Y.-H., Sheu, C.-F., & Wang, W.-C. (2007). Computing confidence intervals of item fit statistics in the family of Rasch models using the bootstrap method. *Journal of Applied Measurement*, *8*, 190–203.
Sykes, R. C., Ito, K., & Wang, Z. (2008). Effects of assigning raters to items. *Educational Measurement: Issues and Practice*, *27*, 47–55.
Taube, K. T. (1997). The incorporation of empirical item difficulty data into the Angoff standard-setting procedure. *Evaluation and the Health Professions*, *20*, 479–498.
Tennant, A. (2004). Disordered thresholds: An example from the Functional Independence Measure. *Rasch Measurement Transactions*, *17*, 945–948.
Tennant, A., & Pallant, J. F. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions*, *20*, 1048–1051.
Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini–Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, *27*, 77–83.
Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*, 358–376.
Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). San Diego, CA: Academic Press.
Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, *23*, 411–440.
Van Nijlen, D., & Janssen, R. (2008). Modeling judgments in the Angoff and contrasting-groups method of standard setting.

Journal of Educational Measurement, *45*, 45–63.
Verheggen, M. M., Muijtjens, A. M. M., Van Os, J., & Schuwirth, L. W. T. (2008). Is an Angoff standard an indication of minimal competence of examinees or of judges? *Advances in Health Sciences Education*, *13*, 203–211.
Verhelst, N. D. (2004). Item response theory. In S. Takala (Ed.), *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section G). Strasbourg, France: Council of Europe/Language Policy Division.
Verhelst, N. D., & Verstralen, H. H. F. M. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 89–108). New York: Springer.
Verhelst, N. D., & Verstralen, H. H. F. M. (2008). Some considerations on the partial credit model. *Psicológica*, *29*, 229–254.
von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement: Manifest variable methods.* Mahwah, NJ: Erlbaum.
Wang, W.-C. (2000). The simultaneous factorial analysis of differential item functioning. *Methods of Psychological Research Online*, *5*, 57–76.
Wang, W.-C., & Liu, C.-Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement*, *67*, 583–605.
Wang, W.-C., & Wilson, M. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*, 296–318.
Wang, W.-C., & Wilson, M. (2005b). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126–149.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263–287.
Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145–178.
Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Basingstoke, UK: Palgrave Macmillan.
Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*, 305–335.
Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (pp. 113–133). Stamford, CT: Ablex.
Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*, 283–306.
Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*, 83–106.
Wolfe, E. W. (2000). Equating and item banking with the Rasch model. *Journal of Applied Measurement*, *1*, 409–434.
Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35–51.
Wolfe, E. W. (2008). RBF.sas (Rasch Bootstrap Fit): A SAS macro for estimating critical values for Rasch model fit statistics. *Applied Psychological Measurement*, *32*, 585–586.
Wolfe, E. W. (2009). Item and rater analysis of constructed response items via the multi-faceted Rasch model. *Journal of Applied Measurement*, *10*, 335–347.
Wolfe, E. W., & Dobria, L. (2008). Applications of the multifaceted Rasch model. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 71–85). Los Angeles: Sage.
Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement*, *2*, 256–280.
Wolfe, E. W., Myford, C. M., Engelhard, G., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP English Literature and Composition examination using benchmark essays* (College Board Research Report No. 2007-2). New York: College Board.
Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah, NJ: Erlbaum.
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transactions*, *16*(3), 888.
Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 1–24). Maple Grove, MN: JAM Press.
Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software* [Computer software and manual]. Camberwell, Australia: ACER Press.
Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education/Praeger.
Zegers, F. E. (1991). Coefficients for interrater agreement. *Applied Psychological Measurement*, *15*, 321–333.
Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.