LANGUAGE ASSESSMENT QUARTERLY, *2*(3), 197–221

Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis

Thomas Eckes
University of Hagen

I studied rater effects in the writing and speaking sections of the Test of German as a Foreign Language (TestDaF). Building on the many-facet Rasch measurement methodology, the focus was on rater main effects as well as 2- and 3-way interactions between raters and the other facets involved, that is, examinees, rating criteria (in the writing section), and tasks (in the speaking section). Another goal was to investigate differential rater functioning related to examinee gender. Results showed that raters (a) differed strongly in the severity with which they rated examinees; (b) were fairly consistent in their overall ratings; (c) were substantially less consistent in relation to rating criteria (or speaking tasks, respectively) than in relation to examinees; and (d) as a group, were not subject to gender bias. These findings have implications for controlling and assuring the psychometric quality of the TestDaF rater-mediated assessment system.

Rater effects such as severity or leniency, halo, or central tendency are commonly viewed as a source of method variance, that is, as a source of systematic variance in observed ratings that is associated with the raters and not with the ratees (Cronbach, 1995; Hoyt, 2000; Myford & Wolfe, 2003). In other words, rater effects are irrelevant to the construct being assessed through ratings and, thus, threaten the validity of the assessment procedure (Bachman, 2004; Messick, 1989, 1995; Weir, 2005).[1]

Correspondence should be sent to Thomas Eckes, TestDaF Institute, University of Hagen, Feithstr. 188, 58084 Hagen, Germany. E-mail: thomas.eckes@testdaf.de

[1]In this article, I am using the term *rater effects* as a generic expression covering various forms of rater errors, inaccuracies, and biases that can be detected as patterns in the ratings assigned to ratees.

In this research, I examined rater effects in *German-as-a-foreign-language* (GFL) performance assessments. GFL assessments of writing and speaking performance are routinely carried out in a standardized fashion using the Test of German as a Foreign Language (TestDaF, *Test Deutsch als Fremdsprache*). Whereas the issue of rater effects has been thoroughly studied in previous research focusing on assessments of English language proficiency (see, e.g., Bachman, Lynch, & Mason, 1995; Engelhard, 1994; Lumley & McNamara, 1995; North, 2000; Weigle, 1999), similar investigations of rater-mediated performance assessments in a GFL context have been lacking.

Filling this gap is important for at least two reasons: (a) the TestDaF is a *large-scale* assessment instrument with many thousands of examinees taking the test worldwide every year, and (b) the TestDaF is a *high-stakes* test designed for foreign students applying for entry to an institution of higher education in Germany. Because of its relevance to far-reaching educational decisions, ensuring a sufficiently high psychometric quality of this measure of German language proficiency is mandatory (see Eckes, 2003, 2005a, 2005b). As part of the quality assurance process, I investigated the extent to which TestDaF writing and speaking performance scores were influenced by differences in rater severity or leniency and by various forms of interaction effects related to raters, such as interactions between raters and examinees or between raters and speaking tasks.

RATER EFFECTS IN LANGUAGE PERFORMANCE ASSESSMENTS

Previous research in various performance settings has revealed substantial degrees of rater effects. Based on different kinds of psychometric approaches implying different ways of modeling these effects, in particular based on generalizability theory (see, e.g., Brennan, 2001; Marcoulides, 2000; Shavelson & Webb, 1991) and many-facet Rasch measurement (Linacre, 1989; Linacre & Wright, 2002), researchers demonstrated the pervasive and often subtle ways in which raters exert influence on ratings. For example, in a meta-analysis of 79 generalizability studies covering a wide range of attribute types and rating procedures, Hoyt and Kerns (1999) found that an average of 37% of variance in ratings was attributable to rater main effects and rater–ratee interactions. When the analysis was restricted to attributes requiring rater inference (e.g., global ratings of achievement or trait ratings), the average proportion of variance due to rater effects was as high as 49%.

Studies focusing on language performance assessments similarly identified a significant degree of rater main effects. In particular, researchers observed marked differences in rater severity or leniency (see, e.g., Engelhard, 1994; Engelhard & Myford, 2003; Lumley & McNamara, 1995). Such differences were shown to exist even after specific rater training (Barrett, 2001; Lumley & McNamara, 1995;

Weigle, 1998) and to persist in groups of raters across a 3-year period (Fitzpatrick, Ercikan, Yen, & Ferrara, 1998). Moreover, researchers found significant effects for rater–ratee interaction (Kondo-Brown, 2002; Lynch & McNamara, 1998), rater–task type interaction (Lynch & McNamara, 1998; Wigglesworth, 1993), and rater–criteria interaction (Wigglesworth, 1993).

Reviewing the implications 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)

In the case of TestDaF, several measures are routinely taken to diminish the degree of unwanted rater variability at the training stage. Thus, raters are trained on the basis of a detailed list of carefully devised scoring guidelines, they are certified to participate in the actual scoring sessions upon fulfillment of strict selection criteria, and they are monitored as to their compliance with scoring specifications on a regular basis. The extent to which these training, evaluation, and monitoring procedures have been successful is an open question—a question that is addressed in this research.

Furthermore, in light of social-psychological research documenting peoples' tendency to see others' behavior through a gender lens (see, for reviews, Deaux & LaFrance, 1998; Eagly & Mladinic, 1994), another issue of importance refers to the extent to which raters are subject to gender-based perceptions and evaluations when scoring examinee performance. This is the issue of *differential facet functioning* related to raters, or *differential rater functioning* (DRF), for short. Note that DRF is conceptually similar to differential item functioning observed across relevant subgroups of examinees (Engelhard & Myford, 2003; Wang, 2000).

Engelhard and Myford (2003) studied DRF in the context of essays written for the Advanced Placement English Literature and Composition Exam. Although, as a group, raters did not show a differential severity or leniency effect related to student gender, the researchers were able to identify individual raters who tended to consistently assign unexpectedly low or high scores to male students' essays, given the particular rater's level of severity or leniency and the male students' performance measures.[2]

In the remainder of this article, I first describe key features of the TestDaF and, then, present results on the psychometric quality of various parts of the TestDaF performance assessment system. Major results concern the degree of differences in

[2]Engelhard and Myford (2003) also studied DRF related to examinee ethnicity and to examinee best language, respectively. In the research reported here, I did not look at these and other background variables in any detail because of too small a database.

rater severity or leniency, the degree of within-rater consistency, as well as two- and three-way interactions between raters and examinees, rating criteria (in the writing section), and tasks (in the speaking section), respectively. Finally, I address the issue of differential rater functioning related to examinee gender.

THE TESTDAF

The data set analyzed here came from the third worldwide administration of the TestDaF that took place in April 2002. As its main purpose, this test is to allow foreign applicants to prove their knowledge of German in the academic context while still in their home country. Test tasks and items are centrally constructed and evaluated, and TestDaF examinee performance is centrally scored (see, for more detail, Eckes et al., 2005; Grotjahn, 2004; see also www.testdaf.de).

The TestDaF consists of four sections: reading comprehension, listening comprehension, written expression, and oral expression. Examinee performance in each of these sections is related to one of three levels of language proficiency in the form of band descriptions; these levels (*TestDaF-Niveaustufen*, TestDaF levels) are *TDN 3*, *TDN 4*, and *TDN 5*. The TDNs are intended to cover the Council of Europe's (2001) Lower Vantage Level (B2.1) to Higher Effective Operational Proficiency (C1.2); that is, the TestDaF measures German language proficiency at an intermediate to high level. There is no differentiation among lower proficiency levels; it is just noted that the TDN 3 level has not yet been achieved (*below TDN 3*).

The writing and speaking sections, which were the focus of this research, are performance-based instruments. More specifically, the writing section is designed to assess the examinees' ability to produce a coherent and well-structured text on a given topic taken from the academic context. In the first part of this section, charts, tables, or diagrams are provided along with a short introductory text, and the examinee is asked to describe the relevant information. Specific points to be dealt with are stated in the rubric. In the second part, the examinee has to consider different positions on an aspect of the topic and write a well-structured argument. The input consists of short statements, questions, or quotes. As before, aspects to be dealt with in the argumentation are stated in the rubric. Both parts are considered tightly interconnected components of a single complex task, and are rated as such.

In a similar fashion, the speaking section taps the examinees' ability to communicate appropriately in typical situations of university life. It consists of 10 tasks and utilizes the Simulated Oral Proficiency Interview format (see Kenyon, 2000). Following this format, the speaking section is administered via audio-recording equipment using prerecorded prompts and printed test booklets. There are four parts, which together comprise 10 tasks. In the first part, the "warm-up task," the examinee is asked to make a simple request. The second part (4 tasks) focuses on situation-related communication, such as obtaining and supplying information,

making an urgent request, and convincing someone of something. The third part (2 tasks) relates to the act of "describing," whereas the fourth part (3 tasks) centers on "presenting arguments."[3] Each task is designed to represent a particular difficulty level. Specifically, at each of the three TDN levels (i.e., TDN 3, TDN 4, and TDN 5) there are three tasks (excluding the warm-up), and the maximum TDN score achievable at a given task corresponds to the difficulty level of that task.[4]

RESEARCH QUESTIONS

The main questions addressed in this research can be summarized as follows:

1. Do TestDaF raters differ in the severity or leniency with which they rate examinee performance in the writing and speaking sections, respectively; and, if so, to which extent?
2. Do TestDaF raters use the TDN rating scale consistently overall; that is, do they stay reasonably close to their own scoring standard?
3. Do TestDaF raters maintain a uniform level of severity or leniency across examinees, across criteria used in the writing section, and across tasks included in the speaking section?
4. Do TestDaF raters show evidence of differential rater functioning related to examinee gender; that is, do they maintain a uniform level of severity or leniency across male and female examinees?

METHOD

Examinees

The writing section was administered to 1,359 participants (747 females, 612 males), the speaking section to 1,348 participants (741 females, 607 males). Par-

[3]Complete sample writing and speaking sections are available online at the following address: www.testdaf.de/html/vorbereitung/modellsatz/ (just add to this address /sa/ for writing, or /ma/ for speaking). Scoring guidelines can be found at www.testdaf.de/html/pruefung/bewertung.htm. Note that since spring 2005, a substantially revised speaking section has been employed (see www.testdaf.de/html/vorbereitung/modellsatz_02/ma/). This new speaking section comprises seven tasks (the warm-up plus 2 tasks per TDN level); moreover, language performance is rated on eight criteria per task, and these ratings are fed into many-facet Rasch analyses comprising four facets (examinees, raters, tasks, and criteria).

[4]The rationale underlying the systematic variation in task difficulty level is, in a sense, akin to the notion of tailored testing. For example, a fairly complex TDN 5 task would demand too much of examinees with language proficiency close to TDN 3, possibly making them fail to achieve the appropriate TDN 3 score, whereas the same examinees would have a real chance to succeed when working on a TDN 3 task.

ticipants' mean age was 23.65 years (*SD* = 4.99); 91.4% of participants were aged between 18 and 30 years.

There were 108 TestDaF test centers involved in this administration (40 centers in Germany, 68 centers in 41 foreign countries). In terms of the number of examinees, the following five national groups ranked highest (percentages in parentheses): the People's Republic of China (30.8%), Bulgaria (20.7%), Russia (5.1%), Morocco (3.0%), and Poland (2.4%).

Raters

The raters who scored the examinees' writing or speaking performance were all experienced teachers and specialists in the field of German as a foreign language. Each rater was licensed upon fulfillment of strict selection criteria. As mentioned previously, raters were systematically trained and monitored as to compliance with scoring guidelines.

Twenty-nine raters (23 women, 6 men) participated in the scoring of examinees' writing performance, 31 raters (26 women, 5 men) provided scorings of examinees' speaking performance. Raters' age ranged from 32 to 68 years. The number of examinees per rater ranged from 29 to 206 for writing, and from 25 to 154 for speaking.

Procedure

Participants were first presented with the writing section (60 min), followed by the speaking section (30 min). Ratings of examinees' essays were carried out according to a detailed catalogue of performance aspects, including grammatical and lexical correctness, range of grammatical and lexical knowledge, degree of structure, and coherence. Based on these specific aspects, raters provided final scorings on the following three criteria: (a) global impression, (b) treatment of the task, and (c) linguistic realization. On each criterion, examinee performance was scored using the 4-point TDN scale (with categories *below TDN 3*, *TDN 3*, *TDN 4*, *TDN 5*).[5]

Regarding the speaking section, examinee oral responses were recorded, and raters scored each speech sample (excluding the warm-up) on the basis of a detailed catalogue of performance criteria. These included fluency, clarity of speech, prosody and intonation, grammatical and lexical correctness, range of grammatical and lexical knowledge, and degree of structure and coherence.

To reduce the raters' workload in terms of the number of tasks to be rated per examinee, and to make this test section more efficient overall, a top-down rating procedure was employed. That is, the rating procedure started with tasks at the TDN 5 level (remember that each task had a predetermined difficulty level). If the

[5]Since Fall 2004, a revised scoring procedure has been employed, in which ratings on nine more detailed criteria provide the input to the analysis.

performance at hand was scored as TDN 5, the rating terminated; otherwise, tasks at TDN 4 were assessed, whereupon it was decided whether tasks at TDN 3 needed to be rated as well. Hence, the number of tasks considered when rating examinee performance in the speaking section could range from three to nine.

Due to this rating design, three different rating scales were used for speaking: (a) a 4-point scale for tasks at TDN 5 level (with categories *below TDN 3, TDN 3, TDN 4, TDN 5*), (b) a 3-point scale for tasks at TDN 4 level (with categories *below TDN 3, TDN 3, TDN 4*), and (c) a 2-point (dichotomous) scale for tasks at TDN 3 level (with categories *below TDN 3, TDN 3*).

In both the writing and the speaking section, examinee performance was independently scored by two raters. These two raters' scorings served as input to the Rasch analysis computer program described next.[6]

Data Analysis

To answer the research questions outlined previously, I analyzed the rating data by means of the computer program FACETS (Version 3.54; Linacre, 2004), with separate FACETS analyses run on the writing and speaking sections. The program used the ratings that raters awarded to examinees to estimate individual examinee proficiencies, rater severities, criteria or task difficulties, and scale category difficulties.

I modeled the rating scale for each criterion (or task) to have its own category structure; that is, the structure of the rating scale could vary from one criterion (or task) to another. Hence, the specific model implemented in the analyses was a *three-facet partial credit model* (Linacre & Wright, 2002). I centered all facets except the examinee facet and left the convergence criteria at their default values (i.e., the maximum size of the largest marginal score point residual was 0.5 score points, and the maximum size of the largest logit change was 0.01 logits; see, for more detail, Linacre, 2004). The estimation process ceased automatically after 165 iterations for writing and after 254 iterations for speaking.

FACETS calibrates the examinees, raters, criteria (tasks), and the rating scales onto the same equal-interval scale (i.e., the logit scale), creating a single frame of reference for interpreting the results of the analysis. Once the parameters of the model have been estimated, interaction effects, such as the interaction between raters and examinees or between raters and criteria, can be detected by examining the standardized residuals (i.e., standardized differences between the observed and expected ratings). An *interaction analysis* (or *bias analysis*) helps to identify unusual interaction patterns among facet elements, particularly those patterns that point to consistent deviations from what is expected on the basis of the model.

---

[6]In the TestDaF examination considered here, final score reporting was based on the average rating computed across raters and criteria (or tasks).

In addition to studying various interaction effects, I performed a *gender bias analysis* by estimating a writing (or speaking) performance measure for each group of female and male examinees along with separate performance measures for each and every combination of individual raters with the respective gender group (see Engelhard, 2002; Engelhard & Myford, 2003; Myford & Wolfe, 2003). More specifically, FACETS computed a gender bias statistic (*Z* statistic) to test the null hypothesis that there was no gender bias in the data. A statistically significant *Z* statistic would indicate that the particular rater–gender group combination resulted in unexpectedly low or high ratings, given the rater's level of severity and the gender group's level of proficiency. Thus, a rater–examinee gender bias analysis helped to find out whether any of the raters exercised differential severity or leniency, rating men's essays (or women's essays) more severely or leniently than expected, or whether each rater's level of severity or leniency was invariant across gender groups.

RESULTS

Global Model Fit

Overall data–model fit can be assessed by examining the responses that are unexpected given the assumptions of the model. According to Linacre (2004), satisfactory model fit is indicated when about 5% or less of (absolute) standardized residuals are equal or greater than 2, and about 1% or less of (absolute) standardized residuals are equal or greater than 3.

There were 8,154 valid responses (i.e., responses used for estimation of model parameters) included in the analysis for writing. Of these, 419 responses (or 5.1%) were associated with (absolute) standardized residuals equal or greater than 2, and 12 responses (or 0.1%) were associated with (absolute) standardized residuals equal or greater than 3. The results for speaking were as follows. Of 16,276 valid responses, 726 responses (or 4.5%) were associated with (absolute) standardized residuals equal to or greater than 2, and 126 responses (or 0.8%) were associated with (absolute) standardized residuals equal or greater than 3.

Taken together, these findings indicated satisfactory model fit for both writing and speaking. Additional statistics that are similarly suited to assess the fit of the data to the Rasch model (e.g., rater fit statistics) are presented later.

Calibration of Examinees, Raters, Criteria, and Tasks

Writing. Figure 1 displays the variable map representing the calibrations of the examinees, raters, criteria, and the 4-point rating scale as raters used it to score

Logit	Examinee	Rater	Criterion	Rating scale for each criterion		
	High	*Severe*	*Hard*	Crit. 1 (TDN 5)	Crit. 2 (TDN 5)	Crit. 3 (TDN 5)
7						
	.					
	.					
6	*					
	**.					
	**.					
5	**.					
	**.					
	***.					
4	*********				----	
	***.			---		---
	********.					
3	******.					
	*******					
	*****.	23				
2	*********.	05			TDN 4	
	*****.			TDN 4		TDN 4
	********.	12				
1	******.	25				
	*********.	03 11 28	3			
	******.	09 15 17 19 29				
0	*********.	06 10 13 16 18 21 22	2		-----	---
	*******.	01 14		---		
	*******.	04 08 26	1			
-1	*******.	07				
	******.					
	****.	20 24 27				TDN 3
-2	*******	02		TDN 3	TDN 3	
	***.					
	*****.					
-3	****.					
	***.			---		
	***.					---
-4	**.				-----	
	**.					
	*.					
-5	*.					
	**.					
-6						
	.					
	.					
-7						
				(below 3)	(below 3)	(below 3)
	Low	*Lenient*	*Easy*			

FIGURE 1 Variable map from the FACETS (Version 3.54; Linacre, 2004) analysis of the Test of German as a Foreign Language (TestDaF) writing performance data. Note that each star in the second column represents seven examinees, and a dot represents fewer than seven examinees. Each number in the fourth column represents a particular rating criterion (1 = *global impression*, 2 = *treatment of the task*, 3 = *linguistic realization*). The horizontal dashed lines in columns 5 through 7 indicate the category threshold measures.

TABLE 1
Summary Statistics for the Many-Facet Rasch Analysis of the Writing Data

Statistics	*Examinees*[a]	*Raters*	*Criteria*
M measure	0.61	0.00	0.00
M SE	0.83	0.14	0.04
χ^2	12,139.3*	1,836.5*	450.3*
df	1,283	28	2
Separation index	4.45	9.61	16.51
Separation reliability	0.91	0.98	0.99

[a]Examinees with nonextreme scores only.
*$p < .01$.

examinee essays on each criterion. Table 1 provides various summary statistics from the FACETS analysis for the three facets.

As can be seen, the variability across raters in their level of severity was substantial. The rater severity measures showed a 4.26-logit spread, which was about a third of the logit spread observed for examinee proficiency measures. Thus, despite all efforts at achieving high rater agreement during extensive training sessions, the rater severity measures were far from being homogeneous. This was consistently revealed by the separation statistics: (a) the fixed chi-square value was highly significant, indicating that at least two raters did not share the same parameter (after allowing for measurement error), (b) the rater separation index showed that within this group of raters there were about nine-and-a-half statistically distinct strata of severity, and (c) the reliability of rater separation attested to a very high rater disagreement.

Speaking. The variable map representing the calibrations of the examinees, raters, tasks, and the 4-point as well as 3-point rating scales as raters used them to score examinee speaking performance is portrayed in Figure 2.[7] Rasch summary statistics for the three facets are shown in Table 2.

Once again, there was a substantial degree of variability across raters in their respective measures. In this FACETS run, the rater severity measures showed a 2.85-logit spread, which was over a fifth of the logit spread observed for examinee proficiency measures. The separation statistics also clearly attested to the high degree of heterogeneity among this group of raters.

Observed and fair scores. The pronounced differences in rater severity measures shown to exist for the TestDaF writing and speaking sections do not at all

[7]The 2-point (dichotomous) rating scale is not included in Figure 2, because in this particular case the category calibrations (or thresholds) for TDN Level 3 tasks (7–9) are the same as these tasks' difficulty measures.

Logit	Examinee	Rater	Task	Rating scale for each task					
	High	*Severe*	*Difficult*	Task 1 (TDN 5)	Task 2 (TDN 5)	Task 3 (TDN 5)	Task 4 (TDN 4)	Task 5 (TDN 4)	Task 6 (TDN 4)
8									
	.								
7	*.								
	*.								
	**.								
6	**.								
	****.								
	***.								
5	****.								
	****.								
	*****.								
4	*****.								
	***.								
	*****.					-----			
3	*****.			-----	-----				
	****.								
	*****.								
2	******.		2					-----	
	*****.					TDN 4	-----		
	*****.	18 21	1	TDN 4	TDN 4				-----
1	*****.	15 17 20 25 28	3						
	****.	12 13 24							
	****.	06 11 16 31	5						
0	****	05 09	6	-----	-----		TDN 3	TDN 3	TDN 3
	***.	03 07 23				-----			
	**.	01 04 14 19 22 27 29	4						
-1	**.	10 30	7						
	**	02 08 26		TDN 3	TDN 3				-----
	*.		9 8			TDN 3	-----		
-2	*.							-----	
	*.								
	.				-----				
-3	.			-----					
	.					-----			
	.								
-4	.								
	.								
-5									
	.								
-6	.								
				(below 3)	(below 3)	(below 3)	(below 3)	(below 3)	(below 3)
	Low	*Lenient*	*Easy*						

FIGURE 2 Variable map from the FACETS (Version 3.54; Linacre, 2004) analysis of the Test of German as a Foreign Language (TestDaF) speaking performance data. Note that each star in the second column represents 11 examinees, and a dot represents fewer than 11 examinees. Each number in the fourth column represents a particular speaking task (1–3 = TDN level 5 tasks rated on a 4-point scale, 4–6 = TDN level 4 tasks rated on a 3-point scale, 7–9 = TDN level 3 tasks rated on a 2-point scale). The horizontal dashed lines in columns 5 through 10 indicate the category threshold measures for the 4- and 3-point scales, respectively (the thresholds for the 2-point scale coincide with the difficulty measures of Tasks 7–9).

TABLE 2
Summary Statistics for the Many-Facet Rasch Analysis of the Speaking Data

Statistics	*Examinees*[a]	*Raters*	*Tasks*
M measure	2.34	0.00	0.00
M SE	0.63	0.10	0.06
χ^2	16,747.8*	2,178.4*	4,387.1*
df	1,240	30	8
Separation index	5.04	10.87	26.24
Separation reliability	0.93	0.98	1.00

[a]Examinees with nonextreme scores only.
*$p < .01$.

stand out as being unusual when related research on rater-mediated performance assessments is taken into account (see, e.g., Bachman et al., 1995; Engelhard, 1994; Lunz & Wright, 1997; McNamara, 1996; Weigle, 1998). Yet, rater severity or leniency 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 TDN levels awarded to examinees may be affected by even small adjustments for differences in rater severity (see, for a detailed discussion, Myford, Marr, & Linacre, 1996).

In this analysis, consider examinees' observed scores, or "raw scores," computed as the average of ratings across raters and criteria (for writing), in relation to these examinees' adjusted scores, or "fair scores," computed on the basis of many-facet Rasch model parameter estimates (Linacre, 2004). The fair scores compensate for rater severity or leniency differences by computing the expected rating for a particular examinee that would be obtained from a rater with level of severity equal to zero; that is, for a rater who was neither more lenient nor more severe than other raters. Table 3 displays some examples taken from the many-facet Rasch analysis of the writing data.

As can be seen, Examinees 616 and 504 received identical observed scores (4.50), yet their fair scores were clearly different (4.85 and 4.03, respectively). Conversely, Examinees 257 and 813 had identical fair scores (4.65) but widely differing observed scores (4.67 and 4.33, respectively). Using conventional cut-offs for the assignment of final TDNs (e.g., TDN 4 if the score is within the range of 3.50 and 4.49), Examinee 813 was awarded a TDN 4 based on the observed score but a TDN 5 (the highest proficiency level possible) based on the adjusted score; exactly the opposite TDN assignments resulted for Examinee 504.

Taking into account the severity measures obtained for the raters involved, the discrepancies between observed and fair scores can be accounted for easily. Thus, Examinee 813's essay was scored by Raters 23 and 19, the first of which proved to be the most severe rater of all raters in the group (2.17 logits, $SE = 0.13$), the other

TABLE 3
Illustrative Results From the Many-Facet Rasch Analysis (Writing Section)

Examinee	*Logit*	*SE*	*Infit*	*Outfit*	*Observed Score*[a]	*Fair Score*[b]	*Number of Ratings*
616	5.43	0.83	0.59	0.58	4.50	4.85	6
257	4.35	0.87	1.18	1.28	4.67	4.65	6
813	4.35	0.83	1.17	1.25	4.33	4.65	6
504	1.92	0.81	0.77	0.76	4.50	4.03	6

Note. Infit and outfit are mean-square fit statistics.

[a]Observed scores are averages computed on the basis of two independent raters' TestDaF-Niveaustufen scores awarded to an examinee's essay on each of three criteria. [b] Fair (or adjusted) scores are averages computed on the basis of model parameters estimated in the FACETS (Linacre, 2004) analysis.

one's severity was still above average (0.26 logits, $SE = 0.12$). Quite to the contrary, both raters involved in the scoring of Examinee 504's essay turned out to be highly lenient (Rater 20: –1.67 logits, $SE = 0.08$; Rater 02: –2.08 logits, $SE = 0.12$; see also Figure 1). In the FACETS run, these severity or leniency differences were controlled for, leading to an upward adjustment of one TDN level for Examinee 813, and to a downward adjustment of one TDN level for Examinee 504, as compared to the TDN assignments based on the observed scores.

TDN assignments across all examinees using fair as compared to observed scores yielded 183 differences (or 13.5% of the sample) in the writing section; more specifically, 140 examinees would have received one TDN level lower than the original score if score adjustment had been employed, whereas 43 examinees would have received one TDN level higher. For speaking, the comparison between fair and observed scores yielded 231 assignment differences (or 17.1% of the sample); in this case, 23 examinees would have received a downward adjustment and 208 examinees an upward adjustment by one level each if the TDN assignment had been based on fair scores. The weighted kappa index (Cohen, 1968) indicated that the agreement between TDN assignments on the basis of fair and observed scores was fairly high overall (.85 for writing, .78 for speaking).

Rater Fit

In this analysis, rater fit refers to the extent to which a given rater is associated with unexpected ratings, summarized over examinees and criteria (in the writing section), or summarized over examinees and tasks (in the speaking section). FACETS reports two mean-square statistics indicating data–model fit for each rater, *rater infit* and *rater outfit.* Whereas the infit statistic is sensitive to an accumulation of unexpected ratings, the outfit statistic is sensitive to individual unexpected ratings. Both statistics have an expected value of 1 and can range from 0 to infinity (Linacre, 2002; Myford & Wolfe, 2003).

TABLE 4
Percentages of Rater Mean-Square Fit Statistics

	Writing		*Speaking*	
Fit Range	*Infit*	*Outfit*	*Infit*	*Outfit*
Narrow				
fit < 0.70 (overfit)	24.1	27.6	0.0	0.0
0.70 ≤ fit ≤ 1.30	75.9	72.4	96.8	90.3
fit > 1.30 (misfit)	0.0	0.0	3.2	9.7
Wide				
fit < 0.50 (overfit)	0.0	0.0	0.0	0.0
0.50 ≤ fit ≤ 1.50	100	100	100	90.3
fit > 1.50 (misfit)	0.0	0.0	0.0	9.7

Note. Infit and outfit are mean-square fit statistics.

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 in their ratings; data provided by these raters tend to *overfit* the model. As a rule of thumb, Linacre (2002) suggested to use 0.50 as a lower control limit and 1.50 as an upper control limit for infit and outfit mean-square statistics.[8] Others researchers suggested to use 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, 2001; McNamara, 1996).

Table 4 presents percentages of rater fit values falling into the overfit, acceptable fit, or misfit categories, using either a narrow or a wide range of upper and lower control limits. Regarding the writing section, there were several raters showing overfit when the fit diagnosis was based on the narrowly defined fit range, whereas none of the raters fell into the misfit category. The overfitting raters had muted ratings that suggested a central tendency or, alternatively, a halo effect (see Engelhard, 2002; Myford & Wolfe, 2004). When the wide fit range was used, all raters showed acceptable fit. The only instances of misfit concerned the speaking section; however, the percentage of raters showing acceptable fit did not fall below 90%, no matter whether the range of fit statistics was set wide or narrow. For the most part, then, raters were internally consistent and used the TDN rating scale appropriately.

Psychometric Dimensionality of the Ratings

Indexes of fit were also used to address the issue of possible psychometric multidimensionality (Henning, 1992; McNamara, 1996) in the writing and speak-

[8]According to Linacre (2002), *outfit* values falling outside the 0.5 to 1.5 fit range are less of a threat to measurement than exceedingly large (or small) *infit* values. Moreover, *misfit* is generally deemed to be more problematic than *overfit* (Myford & Wolfe, 2003).

ing data sets, respectively. Regarding the writing section, the question asked was whether ratings on one criterion followed a pattern that was markedly different from ratings on the others, indicating that examinee scores related to different dimensions, or whether the ratings on one criterion corresponded well to ratings on the other criteria, indicating unidimensionality of the data. The infit values provided by the FACETS analysis were 0.93 (global impression), 1.12 (treatment of the task), and 0.94 (linguistic realization).

Using the same fit-based approach, the dimensionality of the ratings in the speaking section was studied. Infit values for the nine tasks ranged from 0.83 (Task 4) to 1.06 (Task 2).

Because all the indexes were within even narrow quality control limits of 0.70 and 1.30, there appeared to be no evidence of psychometric multidimensionality in either the writing or the speaking data set.

Criterion and Task Discrimination

To investigate whether the criteria used in the writing section and the tasks considered for rating in the speaking section, respectively, were equally discriminating, that is, differentiated equally well between high and low proficiency examinees, the rating scale category calibrations for criteria (or tasks) were examined.[9] Tables 5 to 7 show the category calibrations, or thresholds, for each criterion and each task, respectively, as well as the means and standard deviations of the threshold estimates.

Table 5 reveals that the rating scale category calibrations were fairly consistent across criteria. As Tables 6 and 7 show, the same was true of the category calibrations for the speaking tasks. 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, and for each speaking task alike, were widely separated along the examinee proficiency scale. Thus, on each criterion (and on each speaking task, respectively), examinees had a high probability of being correctly classified into a rating scale category that best described their ability.

Interaction Analysis

Two- and three-way interactions. To investigate whether each rater maintained a uniform level of severity across examinees, or whether particular raters

[9]In a partial credit analysis, the rating scale for each criterion (or task) is modeled to have its own category structure. A rating scale category calibration (or threshold; in FACETS reported as "step calibration") is the point on the examinee proficiency scale at which the probability curves for adjacent categories intersect. On the basis of the 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 could be employed (see Embretson & Reise, 2000; Muraki, 1992; Rost, 2004).

TABLE 5
Rating Scale Category Calibrations for Criteria Used in the Writing Section

	Global Impression		*Treatment of the Task*		*Linguistic Realization*		*Threshold*	
Category	*Threshold*	*SE*	*Threshold*	*SE*	*Threshold*	*SE*	*M*	*SD*
TDN 3	–3.32	0.10	–3.97	0.09	–3.53	0.08	–3.61	0.27
TDN 4	–0.36	0.06	–0.04	0.06	0.02	0.06	–0.13	0.17
TDN 5	3.68	0.07	4.00	0.07	3.51	0.07	3.73	0.20

Note. TDN = TestDaF-Niveaustufen.Thresholds are points on the logit scale at which the probability curves for adjacent categories intersect.

TABLE 6
Rating Scale Category Calibrations for TDN 5-Level Tasks Used in the Speaking Section

	Task 1		*Task 2*		*Task 3*		*Threshold*	
Category	*Threshold*	*SE*	*Threshold*	*SE*	*Threshold*	*SE*	*M*	*SD*
TDN 3	–2.77	0.09	–2.59	0.08	–3.11	0.10	–2.82	0.22
TDN 4	–0.15	0.06	–0.17	0.06	–0.25	0.06	–0.19	0.04
TDN 5	2.91	0.06	2.76	0.07	3.37	0.06	3.01	0.26

Note. TDN = TestDaF-Niveaustufen. Thresholds are points on the logit scale at which the probability curves for adjacent categories intersect.

TABLE 7
Rating Scale Category Calibrations for TDN 4-Level Tasks Used in the Speaking Section

	Task 4		*Task 5*		*Task 6*		*Threshold*	
Category	*Threshold*	*SE*	*Threshold*	*SE*	*Threshold*	*SE*	*M*	*SD*
TDN 3	–1.42	0.11	–1.76	0.08	–1.27	0.09	–1.48	0.20
TDN 4	1.42	0.06	1.76	0.06	1.27	0.06	1.48	0.20

Note. TDN = TestDaF-Niveaustufen. Thresholds are points on the logit scale at which the probability curves for adjacent categories intersect.

scored some examinees' written or oral performance more harshly or leniently than expected, I performed two-way interaction (i.e., Rater × Examinee) analyses. Similarly, I ran a Rater × Criterion, or Rater × Task, interaction analysis to test for patterns of unexpected ratings related to particular criteria used in the writing section, or to particular tasks in the speaking section. Finally, I conducted a three-way interaction analysis to examine whether the combination of a particular rater and a particular criterion, or task, resulted in too harsh or too lenient scores awarded to some examinees.

TABLE 8
Summary Statistics for the Interaction Analysis

	Type of Interaction					
	Rater × Examinee		*Rater × Criterion (or Task)*		*Rater × Examinee × Criterion (or Task)*	
Statistics	*Writing*	*Speaking*	*Writing*	*Speaking*	*Writing*	*Speaking*
N combinations	2,568	2,481	87	279	7,704	15,522
% large *Z* scores[a]	3.6	2.6	37.9	15.4	1.5	1.1
Minimum *Z*	–3.63	–3.08	–4.66	–5.11	–3.08	–3.15
Maximum *Z*	3.46	2.83	4.93	4.47	3.51	3.65
M	0.02	0.02	–0.01	0.02	0.04	0.09
SD	0.94	0.90	2.08	1.54	0.78	0.66

[a]Percentage of absolute *Z* scores (standardized bias scores) equal to or greater than 2.

Table 8 lists the total number of combinations of facet elements considered in each interaction analysis, the percentage of (absolute) *Z* scores equal or greater than 2, the minimum and maximum *Z* scores, as well as their means and standard deviations. Whereas the percentage values for the Rater × Examinee and Rater × Examinee × Criterion (or task) interactions were generally fairly low, more than a third of the combinations of raters and criteria were associated with substantial differences between observed and expected ratings. When raters' scoring behavior was studied in relation to speaking tasks, the percentage of residuals flagged as unexpected was considerably lower, as compared to criteria; yet, still about a sixth of the combinations yielded statistically significant *Z* scores.

Gender bias. The foregoing analysis revealed that only a fairly small percentage of Rater × Examinee interactions produced unexpected responses, regardless of the TestDaF section considered. Hence, there was not much room left for identifying ratings that systematically varied as a function of examinee gender. Still, the question of whether, and to which degree, individual raters were subject to gender bias when scoring essays or audio-recorded utterances remained to be answered, given the potentially harmful consequences for examinees that such a bias can have, how ever small that bias may be. Therefore, I included parameters for the gender facet and the Rater × Examinee gender interaction in the measurement model. This extension of the basic model allowed to study differential rater functioning related to examinee gender.[10]

[10]In early TestDaF examinations as in this one, each essay (in the writing section) and each cassette (in the speaking section), 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. This was to make sure that testing materials were correctly assigned to examinees throughout the testing and scoring process. Following the implementation of automated scanning procedures this practice has been changed, and examinees are now identified by number only.

At the *group-level* analysis, three statistical indicators provided information on potential gender bias: (a) the fixed chi-square statistic helped to find out whether female and male examinees shared the same calibrated level of performance, (b) the gender separation index yielded the number of statistically distinct levels of performance among the gender groups, and (c) the reliability of gender separation showed how well female and male examinees were separated in terms of their performances.

However, as Myford and Wolfe (2004) emphasized, the information provided by each of these summary statistics may be interpreted as demonstrating group-level rater differential severity or leniency only if the researcher has *prior* knowledge about whether the average measures of the gender groups should differ. Because gender differences in verbal ability have been extensively studied, such knowledge was available here (see, e.g., Du & Wright 1997; Engelhard, Gordon, & Gabrielson, 1991; Halpern, 2000; Hyde & Linn, 1988; Johnson, 1996). 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.[11] Thus, the expectation in this study was that women would outperform men, both in the writing and the speaking section, albeit only to a small degree. Evidence of gender bias, therefore, would require that the calibration values for the gender facet were either very small (and not significantly different), indicating gender bias favoring men, or very large (and significantly different), indicating gender bias favoring women.

Table 9 (upper part) provides the relevant summary statistics based on the group-level analysis.

The female–male difference between overall writing measures was 0.44 logits. Women proved to be more proficient than men. As indicated by the chi-square value, the gender difference was statistically significant; both the gender separation index and the separation reliability confirmed this (see, for highly similar findings in a large-scale writing assessment context, Du & Wright, 1997).

Included in Table 9 are also the results of the *individual-level* analysis, indicating whether there were individual raters that displayed differential severity in their ratings. To identify such raters, a bias analysis was performed in which a Rater × Gender group interaction bias term was estimated. FACETS provided two kinds of relevant evidence, each referring to the same underlying bias or interaction information, yet from a different perspective. First, each rater was crossed with each gender group to pinpoint ratings that were highly unexpected given the pattern revealed in the overall analysis ("residual analysis" in Table 9). Second, the severity

[11]The effect size computed for each study was defined as the mean for women minus the mean for men, divided by the pooled within-gender standard deviation (see Hedges & Olkin, 1985).

TABLE 9
Group- and Individual-Level Analysis of Differential Rater Functioning Related to Examinee Gender

Statistics	*Writing*	*Speaking*
Group level		
Women		
Measure	0.22	0.10
SE	0.03	0.02
Men		
Measure	–0.22	–0.10
SE	0.03	0.02
χ^2 (*df*)	89.3* (1)	40.9* (1)
Gender separation index	9.15	6.21
Separation reliability	0.98	0.95
Individual level		
Residual analysis		
% large *Z* scores[a]	1.7	1.6
χ^2 (*df*)	38.8 (58)	65.1 (62)
Pairwise analysis		
% large *t* values[b]	10.3	19.4

[a]Percentage of absolute *Z* scores (standardized bias scores) of Rater × Gender group pairs equal or greater than 2. [b]Percentage of men–women pairs per rater with (absolute) *t* values equal or greater than 2.

*$p < .01$.

of a particular rater when rating women was compared to this rater's severity when rating men, using an approximate *t* test ("pairwise analysis" in Table 9).

Thus, in the writing section, there was only 1 rater–examinee gender combination (out of 58 combinations) that was associated with an unexpectedly high *Z* score (bias measure = 0.38 logits, $SE = 0.19$, $Z = 2.06$). The positive sign of the bias measure (or *Z* score) indicated that this (male) rater on average awarded male examinees lower scores than expected on the basis of the model; the opposite tendency to award female examinees higher scores than expected failed to reach significance (bias measure = –0.22 logits, $SE = 0.14$, $Z = –1.57$). In addition, the overall test that the interaction effects differed from zero was not significant.

Regarding the pairwise analysis, 3 comparisons (or 10.3%) yielded significant *t* values; that is, 2 (male) raters were more severe with male examinees than with female examinees, and 1 (female) rater was more lenient with male examinees than with female examinees. However, when *multiple* comparisons of raters are made (as in the pairwise analyses 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 can be used (see Myers & Well, 2003). Adopting a

conservative approach like this, none of the *Z* or *t* values observed at the individual level of analysis proved to be statistically significant.

Looking at the speaking section, much the same pattern of results showed up, with the difference in measures for women and men amounting to 0.20 logits (see Table 9). Thus, there was a somewhat smaller, yet still significant difference in overall speaking proficiency measures between women and men, with women outperforming men.

At the individual level, only 1 comparison (out of 62 comparisons) yielded a statistically significant *Z* score (bias measure = 0.20 logits, $SE = 0.10$, $Z = 2.10$). As indicated by the bias measure's positive sign, this (female) rater on average awarded male examinees lower scores than expected on the basis of the model; the opposite tendency to award female examinees higher scores than expected failed to reach significance (bias measure = –0.15 logits, $SE = 0.08$, $Z = -1.83$). Just as for writing, the overall test that the interaction effects differed from zero was not significant.

Turning to the pairwise analysis, 6 comparisons (or 19.4%) yielded significant *t* values; that is, 3 (female) raters were more severe with male examinees than with female examinees, and 3 (female) raters were more lenient with male examinees than with female examinees. Yet, again, when a Bonferroni adjustment was used, none of the *Z* or *t* values observed at the individual level of analysis proved to be statistically significant.

SUMMARY AND DISCUSSION

The main purpose of this research was to investigate rater effects in the writing and speaking parts of the TestDaF. Building on the many-facet Rasch measurement approach, I studied rater main effects and interactions between raters and examinees, raters and criteria used for scoring in the writing section, and raters and tasks considered for scoring in the speaking section, as well as three-way interactions involving raters, examinees, and criteria or tasks, respectively.

Analyzing rating data from a worldwide TestDaF administration, I found that raters (a) differed strongly in the severity with which they rated examinees; (b) were fairly consistent in their overall ratings; (c) were substantially less consistent in relation to criteria or tasks, respectively, than in relation to examinees; and (d) as a group, did not show gender bias.

The finding that raters did not function interchangeably in any of the scoring sessions agreed well with related research on the degree of severity exercised in language performance assessments (see, for reviews, Engelhard, 2002; McNamara, 1996; Myford & Wolfe, 2003). At the same time, the spread of rater severity or leniency differences observed in this study was such that in quite a number of cases the actual level of proficiency would have come to be underestimated

when two severe raters had happened to score examinee performance, or to be overestimated when two lenient raters had happened to do the scoring. To compensate for this pronounced rater effect, the many-facet Rasch analysis yields for each examinee an expected performance rating from a hypothetical rater with severity or leniency level equal to zero. This fair average (or fair score) can be used to provide a rater-free estimate of examinee proficiency.

The substantial rater severity or leniency differences shown here, and widely documented in subsequent research on rater effects in other TestDaF performance assessments (Eckes, 2003, 2004a, 2004b), has had three important consequences for the design and implementation of scoring operations in the period following: (a) With respect to rater training, more importance is attached to within-rater consistency than to between-rater homogeneity in the use of scoring standards, (b) raters are constantly monitored on the basis of their severity or leniency and consistency, (c) final TDN writing and speaking scores are awarded to examinees on the basis of their fair scores, that is, after adjusting their observed scores for differences in rater severity.

In terms of their overall use of the TDN rating scale, raters appeared to be internally consistent in scoring examinee writing or speaking performance. Yet, they were less consistent when moving from one criterion in the writing section, or task in the speaking section, to the next. The FACETS interaction analysis revealed that about 37% of the Rater × Criterion combinations, and about 16% of the Rater × Task combinations, produced unexpectedly high deviations from model expectations. Because the rating criteria were designed to tap different aspects of examinee writing performance, the fairly high percentage of flagged interactions for writing is less of a problem than it may first seem. Nonetheless, the fact that raters exercised more severity or leniency with some criteria than with others, as compared to model expectations, points to unwanted rater variability in construing the meaning of each criterion. With speaking, the situation is somewhat different, because each task is designed to assess the same underlying dimension, albeit at different levels of proficiency. Hence, inconsistency in ratings across tasks highlights another, maybe even more problematic aspect of the assessment system. As one way to counter these problems, the scoring procedure used in both sections has been revised such as to raise the rater consistency in relation to criteria and tasks, respectively.

In this study, I also tested whether raters displayed a gender bias when scoring examinees' essays or speaking performance. Adding a gender facet and a Rater × Gender group interaction bias term to the basic measurement model allowed to address this issue. The results showed that, on average, women were awarded higher scores than men in both the writing and the speaking sections—a result that was in line with expectations based on prior research on gender differences in verbal ability. The observed difference, therefore, could not be interpreted as evidence of gender bias exercised by the raters as a group. However, in some individ-

ual cases, tendencies to score women more leniently than men, or vice versa, turned up. Though failing to reach the level of significance when a Bonferroni adjustment was used, these biased tendencies were somewhat more pronounced in the speaking than in the writing section, presumably because examinee gender was easily conveyed to raters in the speaking section by vocal features of examinee performance.

It was beyond the scope of this article to give a detailed account of other important kinds of rater effects potentially manifesting themselves in performance assessments (see, for an in-depth discussion, Myford & Wolfe, 2004; Wolfe, 2004). The focus was on severity or leniency and special forms of differential severity or leniency. Data relevant to detecting other effects such as central tendency, randomness, and halo, were not considered in any detail. In particular, the halo effect could be an issue due to the specifics of the scoring procedure used with the TestDaF writing and speaking sections. Work is currently in progress to investigate halo and other kinds of rater effects utilizing the many-facet Rasch measurement framework.

In future research on rater-mediated performance assessments in general, and TestDaF assessments in particular, other important issues to address, or to study more intensely, include the following: (a) What factors account for the differences in severity or leniency that raters exercise when scoring examinee performance? (b) How stable (or variable) are severity measures across different scoring sessions, contexts, and points in time? (c) How do raters deal with feedback on their severity or leniency level and degree of scoring consistency provided by findings from many-facet Rasch analysis? (d) Does differential rater functioning exist in relation to other examinee background variables (e.g., ethnicity or second language)? (e) Do raters fall into types characterized by distinctive and coherent patterns of scoring tendencies, and, if so, what is the nature of these patterns? In the end, answers to questions like these serve to provide examinees with test scores as objective, valid, and fair as possible.

CONCLUSION

Rater effects are a perennial and ubiquitous phenomenon. They usually come in many forms and can hide in many parts of an assessment system. In this research, the many-facet Rasch measurement approach was used to detect potential ramifications of rater effects in assessments of writing and speaking performance in the Test of German as a Foreign Language (TestDaF). Specifically, through fine-grained interaction analyses this approach was able to pinpoint aspects of the TestDaF assessment system that were functioning as intended, as well as potentially problematic aspects. The information thus gained has been guiding the revi-

sion of rater-mediated assessments as an integral part of the quest for further improvement on the TestDaF's psychometric quality.

ACKNOWLEDGMENT

I thank four anonymous reviewers of *Language Assessment Quarterly* for helpful comments on earlier drafts of this article.

REFERENCES

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12,* 238–257.

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

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213–220.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, England: Cambridge University Press.

Cronbach, L. J. (1995). Giving method variance its due. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A Festschrift in honor of Donald Fiske* (pp. 145–157). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Deaux, K., & LaFrance, M. (1998). Gender. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 788–827). Boston: McGraw-Hill.

Du, Y., & Wright, B. D. (1997). Effects of student characteristics in a large-scale direct writing assessment. In M. Wilson, G. Engelhard Jr., & K. Draney (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 1–24). Stamford, CT: Ablex.

Eagly, A. H., & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology, 5,* 1–35.

Eckes, T. (2003). Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse [Quality assurance with the TestDaF: Concepts, methods, results]. *Fremdsprachen und Hochschule, 69,* 43–68.

Eckes, T. (2004a). 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 Deutsch als Fremdsprache" (TestDaF)]. *Diagnostica, 50,* 65–77.

Eckes, T. (2004b). Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen [Facets of language testing: Severity and consistency in language performance assessments]. In A. Wolff, T. Ostermann, & C. Chlosta (Eds.), *Integration durch Sprache* (pp. 485–518). Regensburg, Germany: FaDaF.

Eckes, T. (2005a, May). *Assuring the quality of TestDaF examinations.* Paper presented at the ALTE 2nd International Conference, Berlin, Germany.

Eckes, T. (2005b). 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., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., et al. (2005). Progress and problems in reforming public language examinations in Europe: Cameos from the Baltic States, Greece, Hungary, Poland, Slovenia, France, and Germany. *Language Testing, 22,* 355–377.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31,* 93–112.

Engelhard, G., Jr. (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: Lawrence Erlbaum Associates, Inc.

Engelhard, G., Jr., 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., Jr., & 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 Rep. No. 2003–1). New York: College Entrance Examination Board.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11,* 195–208.

Grotjahn, R. (2004). TestDaF: Theoretical basis and empirical research. In M. Milanovic & C. J. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference July 2001* (pp. 189–203). Cambridge, England: Cambridge University Press.

Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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.

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.

Johnson, S. (1996). The contribution of large-scale assessment programmes to research on gender differences. *Educational Research and Evaluation, 2,* 25–49.

Kenyon, D. M. (2000). Tape-mediated oral proficiency testing: Considerations in developing Simulated Oral Proficiency Interviews (SOPIs). In S. Bolton (Ed.), *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests* [TESTDAF: Foundations of developing a new language test] (pp. 87–106). München, Germany: Goethe-Institut.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19,* 3–31.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago: MESA Press.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2004). *A user's guide to FACETS: Rasch-model computer programs* [Software manual]. Chicago: Winsteps.com

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3,* 484–509.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12,* 54–71.

Lunz, M. E., & Wright, B. D. (1997). Latent trait models for performance examinations. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 80–88). Münster, Germany: Waxmann.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15,* 158–180.

Marcoulides, G. A. (2000). Generalizability theory. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 527–551). San Diego, CA: Academic.

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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 Rep. No. 95–40). Princeton, NJ: Educational Testing Service.

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.

Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion* (2nd ed.) [Textbook test theory, test construction]. Bern, Germany: Huber.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Wang, W.-C. (2000). The simultaneous factorial analysis of differential item functioning. *Methods of Psychological Research Online, 5,* 57–76.

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.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Houndmills, England: Palgrave Macmillan.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing, 10,* 305–335.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46,* 35–51.