# Rater-Mediated Listening Assessment: A Facets Modeling Approach to the Analysis of Raters' Severity and Accuracy When Scoring Responses to Short-Answer Questions

Thomas Eckes[1]

[1] TestDaF Institute, University of Bochum

**Abstract:**
Short-answer questions are a popular item format in listening tests. Examinees listen to spoken input and demonstrate comprehension by responding to questions about the information contained in the input. Usually, human raters or markers score examinee responses as correct or incorrect following a scoring guide. Considering this procedure an instance of the more general class of rater-mediated language assessment, the present research adopted a many-facet Rasch measurement approach to provide a detailed look at the psychometric quality of the listening scores. Nine operational raters and one expert rater scored responses of 200 examinees to 15 short-answer questions included in the listening section of a standardized language test. The findings revealed that (a) raters differed significantly in their severity measures, albeit to a lesser extent than typically observed in writing or speaking assessments, (b) raters did not show evidence of differential severity across short-answer questions, (c) raters evidenced an overall high level of scoring accuracy, but also showed non-negligible differences in their accuracy measures, and (d) raters did not show evidence of differential accuracy across short-answer questions. Implications for the validity and fairness of using short-answer questions in listening tests as well as for rater training and monitoring purposes are discussed.

**Keywords:**
*rater-mediated assessment, listening assessment, short-answer questions, facets models, rater severity, rater accuracy*

## Introduction

Language tests and assessments play an increasingly important role in educational, employment, and immigration contexts (Kunnan, 2012; McNamara et al., 2019). Based on language test results, different kinds of high-stakes decisions are taken in contemporary society. For example, in higher education where international students apply for entry to academic programs, admissions decisions critically hinge on the scores achieved on language proficiency tests (Eckes & Althaus, 2020; Elder, 2017). These tests measure the applicants' linguistic preparedness for study in a language other than their best or native language. Notably, the tests provide information about the proficiency level reached in each of the four language skills of reading, listening, writing, and speaking. The focus of the present research is on the assessment of listening skills.

The assessment of listening faces many specific challenges. Compared to the productive skills of writing and speaking, listening is an invisible process that does not entail directly observable output. Different from reading, listening occurs in real-time. Listeners cannot refer back to the spoken input as readers usually can do with written input. When the listening event is over, what remains is a transient, more or less fragmentary representation of the spoken input in the listener's mind. Also, the exact content of that representation is affected by a wide range of variables, most of which refer to the speaker, the spoken input, the listening context, and the listener (Buck, 2001; Green, 2017; Rost, 2016).

In a standard listening assessment, examinees are first required to listen to a spoken input, for example, a lecture, a dialogue, or a group discussion, and then to provide evidence of comprehension. When creating listening tasks, one of the many issues that need to be addressed concerns the kind of input to present to listeners. For example, test developers need to decide whether the input should be authentic, that is, closely related to real-world spoken language, or scripted (Ockey & Wagner, 2018; Rost, 2016). Another question is how listeners are expected to respond to the input. Test developers have to choose an item format that is suited to support the intended inferences about the examinees' listening ability beyond the assessment context.

Item formats in widespread use in listening tests include multiple-choice items, multiple-matching tasks, and short-answer questions (Field, 2019; Green, 2017; Kang et al., 2019). The present study is concerned with the short-answer format – a type of limited production task where examinees have to produce or construct a response (Bachman & Palmer, 2010; Carr, 2011). Specifically, examinees are required to respond to questions that refer to the content of the spoken input by writing down single words or short phrases. Hence, short-answer questions allow test developers to capture the real-world activity of note-taking, which is typical of the academic context (Song, 2011). This is also the context of the listening test under study here.

Usually, human raters, in listening assessments also called (clerical) markers, scorers, or assessors, evaluate the correctness of the examinees' responses building on a scoring (marking) guide. The use of short-answer questions, therefore, bears a striking resemblance to rater-mediated language assessments more generally (Eckes, 2015, 2019; Engelhard & Wind, 2018): The information that an observed response provides about the construct of interest (i.e., listening abili-

ty) is mediated through the rater's judgmental and decision-making processes (Eckes, 2017, 2019; Engelhard et al., 2018).

The present study focuses on raters or, more precisely, on rater characteristics essential for establishing the validity and fairness of the listening assessment outcomes. In particular, adopting a many-facet Rasch measurement approach with raters as a separate facet, this study investigates (a) the severity or leniency of each rater and (b) the accuracy of the scores raters assign to examinee responses. By doing so, the study addresses key demands set forth by Lane (2019): "The use of rater-mediated assessments requires the evaluation of the accuracy and consistency of the inferences made by those who interpret examinee performances to ensure the validity of their judgments regarding examinee performances and the uses of the examinee scores" (p. 653). The study intends explicitly to close a gap in research on the quality of raters' judgments when scoring responses to short-answer questions in listening assessments.

## Implications for scoring examinee listening performances

Short-answer questions are suited to target different skills or components of listening comprehension (Buck, 2001; Green, 2017). These components include lower-level skills, such as phoneme decoding or syntactic parsing, and higher-level skills, such as meaning construction – a particularly appealing feature for tests that aim at assessing a range of different listening ability levels. Short-answer questions are also largely unaffected by guessing strategies (Green, 2017; McNamara, 2000). When combined with selected-response item formats (e.g., multiple-choice or multiple-matching tasks), short-answer questions help to control for item format effects (In'nami & Koizumi, 2009). Because each format places somewhat different processing demands on examinees (Brindley & Slatyer, 2002; Field, 2019), relying on just one item format, for example, multiple-choice items, may put examinees at a disadvantage who would fare better when also allowed to respond to multiple-matching tasks or short-answer questions.

On the downside, the short-answer format comes with its problems (Buck, 2001; Green, 2017). First of all, this format requires (a) availability of specifically trained raters and (b) construction of a detailed scoring guide. Moreover, raters may encounter cases where they are forced to make decisions under uncertainty even when these two requirements are met. Scoring responses to short-answer questions generally involves deciding on whether a given response is correct or incorrect; that is, raters have to make binary decisions. Considering that listening comprehension forms a latent ability continuum, Buck (2001) noted that awarding 1 point for each correct answer and 0 points for each incorrect answer, "turns a complex continuum into a simple dichotomy. The scorer has to make a *whole series of decisions* [emphasis added] about which responses are acceptable, and which are not" (Buck, 2001, p. 141).

Depending on the specific listening skills targeted by short-answer questions, the decisions that raters have to make may threaten the fairness and accuracy of the assigned scores to varying degrees. For example, when targeting listening to identify specific information, examinees frequently write just one word (or a maximum of two words) to demonstrate comprehension. In that case, raters may simply refer to the correct

responses or keywords listed in the scoring guide, awarding 1 point if the response in question is contained in that list and 0 points otherwise. On the other hand, when targeting understanding of main ideas or gist, examinees often write more than two words or short phrases; as a result, responses are much more complex and varied. It is becoming increasingly difficult for raters to decide on the correctness or adequacy of the responses in such cases.

Within the context of assessing writing or speaking, decision-making processes and the closely related issues of rating quality, rater effects, and rater biases have been extensively studied (Eckes, 2017, 2019; Engelhard, 2002; Engelhard & Wind, 2018; McNamara et al., 2019; Wind & Peterson, 2018). By contrast, research on the specific ways that raters deal with the uncertainties of scoring listening performances, as well as research on the implications that these uncertainties have for the validity of the assessment outcomes, has been very sparse. To illustrate, in a review of statistical methods for evaluating rating quality in language assessments covering a total of 259 methodological and applied studies, Wind and Peterson (2018) identified 156 studies focusing on L1 or L2 writing (135 studies) or speaking (21 studies), but no more than one study on listening (using classical rater agreement statistics). Similarly, in a review of Rasch model applications in language assessment from 2000 to 2018, Fan and Knoch (2019) found that out of a total of 64 studies on rater effects, 38 studies were concerned with writing and 20 studies with speaking, but not a single study with listening.

## Context of the present study: The TestDaF listening assessment

The Test of German as a Foreign Language (*Test Deutsch als Fremdsprache*, TestDaF) is officially recognized as a language exam for international students applying for entry to higher education institutions in Germany. The TestDaF assesses the four language skills in separate sections (reading, listening, writing, and speaking). Examinee performance in each section is related to one of three levels of language proficiency, the so-called TestDaF levels (*TestDaF-Niveaus*, TDNs). The levels TDN 3, TDN 4, and TDN 5 cover the Council of Europe's (2001) Lower Vantage Level (B2.1) to Higher Effective Operational Proficiency (C1.2); that is, the test measures German language proficiency at an intermediate to high level. Examinees achieving at least TDN 4 in each section are eligible for admission to a German institution of higher education ( for a review of the TestDaF, see Norris & Drackert, 2018; see also https://www.testdaf.de, where free sample tests are available).

The TestDaF listening section assesses an examinee's ability to understand spoken texts thematically and linguistically related to higher education. It consists of three parts. Each part is based on a scripted listening text, audio-recorded and played to examinees in group sessions. The first part presents a dialogue often encountered at university (e.g., a conversation between two students at the library). The second part presents a radio interview or a discussion on a familiar academic topic involving three or four speakers (e.g., a discussion on requirements for enrolling in academic programs). Finally, the third part presents a short lecture or an interview with an expert focusing on a scientific issue of general interest (e.g., an introductory lecture about climate change).

Examinees listen to the three texts presented in order of increasing difficulty, as roughly defined in terms of three text characteristics: (a) level of abstraction or informational density, (b) complexity of sentence structures, and (c) number of words (e.g., 350 to 400 words for Text 1, 550 to 580 words for Text 2, 580 to 620 words for Text 3). Audio recordings of the first two listening texts are presented only once; the audio recording of the third text is presented twice.

There are two types of listening items. The first type comprises 15 short-answer questions. Eight of these items belong to the first part of the listening section; the remaining seven items belong to the third part. Responses to each of these items are guided by questions asking for specific content conveyed by the speakers involved (e.g., "What are the new opening times of the library?", "What kind of research evidence does the lecturer refer to?"). Examinees are required to listen to each text and respond to the questions while listening by noting the relevant keyword(s) on an answer sheet. The second item type, presented in the second part of the listening section, comprises 10 true/false items. Examinees have 30 minutes to respond to all 25 items and another 10 minutes to transfer their responses to machine-readable answer sheets.[1]

The responses to the short-answer questions are scored by trained raters using a predefined list of correct answers. Responses that are contained in this list are scored 1 point; responses that differ from the correct solution in terms of minor orthographical, grammatical, lexical, or semantical deviations are also scored 1 point. All other responses are scored 0 points.

The listening data analyzed here were collected as part of an ongoing validation program, focusing on the TestDaF writing, speaking, and listening sections. Regarding the listening section, a sample of examinee responses to short-answer questions that formed part of an earlier live exam was presented to a small group of raters for scoring. The procedure differed in two critical ways from the standard TestDaF scoring procedure: First, all raters scored all listening performances, whereas in the standard procedure each examinee's performance is scored only once. The standard single-rater scoring design rests on the commonly-held assumption that raters, when they are experienced and specifically trained, will show negligibly small differences in scoring behavior (if any).[2] The present study investigated this assumption. Second, in addition to the group of operational raters, the validation study's scoring design included an expert rater. The scores assigned by the expert served as a kind of benchmark against which to assess the accuracy of the operational raters. Including a single expert or a group of experts ("validity committee") is quite common in measurement approaches to studying the accuracy or validity of scores assigned to writing performances (e.g., Engelhard, 1994, 1996; Jin & Wang, 2018; Wind & Engelhard, 2013; Wolfe & McVay, 2012). By contrast, this procedure has not yet been implemented within the context of listening assessments.

1 The three parts of the TestDaF listening section form testlets (Wainer & Kiely, 1987), with 8, 10, and 7 items, respectively. A study of this test structure revealed negligibly small testlet effects for the first and third part of the listening section (Eckes, 2014); the moderate testlet effect observed for the second part has no implications for the findings presented here.

2 For an example of listening assessments that expect markers to show machine-like scoring behavior, see Geranpayeh (2013, p. 264).

## Research questions

The questions guiding the present research aimed at two basic quality dimensions of scoring examinee responses to short-answer listening items: (a) the rater severity dimension and (b) the rater accuracy dimension. Both of these dimensions were studied in detail, building on a many-facet Rasch measurement approach. The specific research questions were as follows:

1. Do raters differ in their severity or leniency when scoring responses to short-answer questions in listening assessments, and, if so, how pronounced are these differences?
2. Do raters show evidence of differential rater functioning in terms of varying levels of severity or leniency across short-answer questions?
3. (a) How accurate are the scores; that is, how much do scores provided by the operational raters agree with scores provided by the expert rater? (b) Do the operational raters differ in their accuracy when scoring examinee listening performances, and, if so, how pronounced are these differences?
4. Do raters show evidence of differential rater functioning in terms of varying levels of accuracy across short-answer questions?

## Method

### Overview

Trained raters scored responses of examinees to short-answer questions in the TestDaF listening section. Raters provided dichotomous scores; that is, they marked responses as correct or incorrect, using a detailed scoring guide. The observed scores were subjected to a facets analysis to gain insight into the variability of raters along the severity dimension. In a second approach, accuracy scores were first derived by comparing scores assigned by the operational raters with scores from the expert rater. Then, the accuracy scores were subjected to a facets analysis to gain insight into the variability of raters along the accuracy dimension. Finally, differential rater severity and accuracy effects related to short-answer questions were examined.

### Participants

Nine operational raters and one expert rater marked the responses of 200 examinees to 15 short-answer questions presented as part of the TestDaF listening section. The operational raters (8 females, 1 male) were all experienced teachers or specialists in German as a foreign language. Each rater was licensed upon fulfillment of strict selection criteria and systematically trained and monitored as to compliance with scoring guidelines. The expert rater had been instrumental in supervising the development of listening items, co-authoring the rater manual for the listening section, and selecting, training, and monitoring the operational raters.

This study's listening performances were sampled from the total group of 3,949 examinees taking the TestDaF in April 2012 (2,557 females, 1,392 males). All examinees were international students applying for entry to an institution for higher education in Germany. For each examinee, two kinds of data were available: (a) a listening score, ranging from 0 to 25 score points, and (b) a final TDN level for listening. Based on the listening scores and the TDNs, performances of 200 examinees were selected for the present validation study. Specifically, to cover the range of examinee proficiencies most critical in terms of eligibility for university admission (i.e., TDN levels 3 and 4), 100 examinees were randomly drawn from among examinees who scored near the borderline for each of these levels at the lower end of the TDN scale (i.e., below TDN 3 vs. TDN 3 and TDN 3 vs. TDN 4); another 100 examinees were drawn from the entire group at random. The final sample considered here comprised 138 females and 62 males.

## Procedure

Each of the operational raters and the expert rater scored the responses of all 200 examinees to the short-answer questions; that is, the maximum possible number of scores per rater was 3,000. Owing to missing values, the actual number of scores available for data analysis was slightly lower (the proportion of missing values was 4.52%). The scoring design underlying this study was complete (fully crossed), ensuring strong connectedness between examinees, raters, and items.

## Data analysis

Following the scaled ratings tradition of investigating rating quality (Eckes, 2017; Wind & Peterson, 2018), as mentioned before, this study built on a many-facet Rasch measurement (or facets modeling) approach (Eckes, 2015, 2019; Engelhard, 2013; Engelhard & Wind, 2018). The computer program FACETS (Version 3.81; Linacre, 2018a) was used to analyze the listening data. FACETS uses joint maximum likelihood (JML) estimation of the model parameters (for a discussion of the JML approach, see Linacre, 2018b; Robitzsch & Steinfeld, 2018). Specifically, two different kinds of facets models were applied to provide evidence on the rating quality from a range of perspectives: (1) a severity facets model, and (2) an accuracy facets model.

First, the analysis based on the severity facets model focused on studying differences between raters regarding their tendency to award scores to examinees in a rather severe or lenient manner (Eckes, 2015, 2019; Engelhard, 2013). The input to the rater severity analysis consisted of the set of dichotomous scores provided by the operational raters and the expert rater for the examinee responses to each of the short-answer items.

Second, the analysis based on the accuracy facets model focused on studying the accuracy of the operational raters in terms of their agreement with the expert (Engelhard, 2013; Wind & Engelhard, 2012, 2013; Wolfe et al., 2015). Operational marks that were in exact agreement with the expert marks were considered accurate and assigned an accuracy score of 1, and those that were different were considered inaccurate and assigned an accuracy score of 0. The rater accuracy analysis input consisted of the set of dichotomous accuracy scores computed for each operational rater; greater scores indicated higher rater accuracy.

***Rater severity***

The severity facets model was a three-facet extension of the dichotomous Rasch model (Rasch, 1960/1980). More specifically, this model was a main-effects model given as follows:

$$ln\left[\frac{p(x_{nij}=1)}{p(x_{nij}=0)}\right]=\theta_n-\beta_i-\alpha_j \quad (1)$$

where $x_{nij}$ is the score of examinee $n$ for item $i$ assigned by rater $j$, with $x_{nij} = 1$ for a correct response and $x_{nij} = 0$ for an incorrect response; $\theta_n$ is the ability of examinee $n$; $\beta_i$ is the difficulty of item $i$; and $\alpha_j$ is the severity of rater $j$.

For the rater severity analysis, where the expert rater is conceived of as a reference, the rater facet was anchored by setting the expert's severity measure to 0.0 logits. As usual, the item facet was centered, that is, constrained to have a mean element measure of 0.0 logits; the examinee facet was left non-centered, that is, examinee measures were free to float relative to rater and item measures.

***Rater-by-item interaction***

The severity facets model was modified by adding a rater-by-item interaction parameter for studying differential rater functioning in terms of interactions between raters and items (Eckes, 2015; Engelhard & Wind, 2018; Myford & Wolfe, 2003). The interaction model statement underlying this analysis was:

$$ln\left[\frac{p(x_{nij}=1)}{p(x_{nij}=0)}\right]=\theta_n-\beta_i-\alpha_j-\omega_{ij} \quad (2)$$

where $\omega_{ij}$ is the interaction parameter (also called bias parameter or bias term); all other terms are defined as in Equation 1.

The rater-by-item interaction analysis aimed at investigating whether particular raters scored examinee responses to particular items more severely or more leniently than expected given the raters' severity measures and the items' difficulty measures estimated based on the severity facets model (Eq. 1). According to Rasch model expectations, each rater's severity should be invariant across items.

***Rater accuracy***

Similar to the severity facets model, the accuracy facets model was a three-facet extension of the dichotomous Rasch model (Rasch, 1960/1980) given as follows:

$$ln\left[\frac{p(y_{nij}=1)}{p(y_{nij}=0)}\right]=\delta_n+\sigma_i+\lambda_j \quad (3)$$

where $y_{nij}$ is the accuracy score of operational rater $j$ for examinee $n$'s response to item $i$, with $y_{nij} = 1$ for a match between the operational and the expert mark and $y_{nij} = 0$ for a mismatch; $\delta_n$ is the easiness of providing an accurate mark for responses of examinee $n$; $\sigma_i$ is the easiness of providing an accurate mark for responses to item $i$; and $\lambda_j$ is the accuracy of rater $j$.

Note that the parameters for the elements of all three facets have a positive orientation. In particular, higher values of the rater accuracy parameter mean higher accuracy scores, that is, higher proportions of agreement with expert marks. Also, different from an analysis based on the rater severity model (Eq. 1), the rater accuracy analysis targets the rater facet; that is, raters are the objects of measurement instead of examinees (Wind & Engelhard, 2013). Therefore, in this analysis, the examinee and item facets were centered, that is, constrained to have a mean element measure of 0.0 logits; the rat-

er facet was left non-centered, that is, rater measures were free to float relative to examinee and item measures.

***Rater-by-item interaction***

The accuracy facets model was modified by adding a rater-by-item interaction parameter for studying differential rater functioning in terms of interactions between raters and items (Eckes, 2015; Engelhard et al., 2018; Engelhard & Wind, 2018; Myford & Wolfe, 2003). The interaction model statement then became:

$$ln\left[\frac{p(y_{nij}=1)}{p(y_{nij}=0)}\right] = \delta_n + \sigma_i + \lambda_j + \eta_{ij} \quad (4)$$

where $\eta_{ij}$ is the interaction parameter; all other terms are defined as in Equation 3.

The rater-by-item interaction analysis aimed at investigating whether particular raters scored examinee responses to particular items more accurately or more inaccurately than expected given the raters' accuracy measures and the items' easiness measures estimated based on the rater accuracy model (Eq. 3). Ideally, each rater's accuracy should be invariant across items.

## Results

### Rater severity analysis

***Wright map***

Figure 1 displays the measures for examinee proficiency, rater severity, and item difficulty in a common frame of reference. The variability across examinees in their level of estimated proficiency was substantial. The examinee proficiency measures showed an 8.20-logit spread, which was highly congruent with the typical range of examinee measures in the TestDaF listening section (varying between 8 and 9 logits). Quite in contrast to the examinee facet, the measurement results for the rater facet revealed a greatly reduced variability. The rater locations tightly clustered around 0.0 logits, the value to which the expert's measure was set for anchoring the rater facet. Finally, the item difficulty measures appeared widely separated along the logit scale.

***Summary Rasch statistics***

The summary Rasch statistics shown in Table 1 provide an overview of the variability among the elements of the examinee, rater, and item facets. Importantly, as the rater statistics show, the between-rater severity differences, though relatively small, were nonetheless statistically significant. More specifically, the homogeneity index $Q$ for the rater facet, which is approximately distributed as a chi-square statistic with nine degrees of freedom (*df*), indicates statistically significant differences between severity measures of at least two raters, $Q(9) = 38.0$, $p < .01$. The rater separation (or number of strata) index $H$ confirms that the present sample of raters was separable into more than two and a half distinct levels or classes of severity (perfect rater homogeneity would be indicated by an $H$ value close to 1). Finally, the rater separation reliability value of .76 is just another way to express the fact that the degree of severity differences within the present rater sample was not negligibly small (perfect interchangeability of raters would be indicated by an $R$ value close to 0).

```
+-------------------------------------------------------------------------------+
|Measr|+Examinees  |-Raters                                                     |-Items  |
|-----+------------+------------------------------------------------------------+--------|
|   5 +   High     +                        Severe                              + Diffi- |
|     |            |                                                            | cult   |
|     |            |                                                            |        |
|     | .          |                                                            |        |
|     | .          |                                                            |        |
|   4 +            +                                                            +        |
|     |            |                                                            |        |
|     |            |                                                            |        |
|     |            |                                                            |        |
|     | .          |                                                            |        |
|   3 + *          +                                                            +        |
|     | ***        |                                                            |        |
|     | *          |                                                            |        |
|     | **         |                                                            |        |
|     | **.        |                                                            |        |
|   2 + *.         +                                                            +        |
|     | ****       |                                                            | 22     |
|     | ***.       |                                                            |        |
|     | **.        |                                                            | 19     |
|     | ****.      |                                                            |        |
|   1 + ****.      +                                                            + 20     |
|     | *******    |                                                            |        |
|     | *********. |                                                            | 1   21 |
|     | *****      |                                                            | 25     |
|     | ***        | 1       2       3       4       5       7       9          | 6      |
*   0 * ********   * 6       8       Expert                                     * 24  5  *
|     | *****.     |                                                            |        |
|     | ****       |                                                            | 2   3  |
|     | ******     |                                                            | 7      |
|     | ****.      |                                                            |        |
|  -1 + ****.      +                                                            +        |
|     | ***        |                                                            | 4      |
|     | *.         |                                                            |        |
|     | *.         |                                                            | 8      |
|     | *          |                                                            | 23     |
|  -2 + *.         +                                                            +        |
|     |            |                                                            |        |
|     | *          |                                                            |        |
|     | .          |                                                            |        |
|     | .          |                                                            |        |
|  -3 +            +                                                            +        |
|     | .          |                                                            |        |
|     |            |                                                            |        |
|     |            |                                                            |        |
|     | .          |                                                            |        |
|  -4 +    Low     +                        Lenient                             +  Easy  |
|-----+------------+------------------------------------------------------------+--------|
|Measr| * = 2      |-Raters                                                     |-Items  |
+-------------------------------------------------------------------------------+
```

**Figure 1** Wright map for the rater severity analysis

***Rater severity calibrations***

Table 2 presents the detailed measurement results for the nine operational raters and the expert rater. Since the expert severity measure was anchored at 0.0 logits, it is readily seen that the majority of operational raters assigned scores that tended toward the severe end of the measurement scale. The *SE* (standard error) column indicates that the precision of the severity measures was

**Table 1** Summary Rasch statistics for the many-facet rater severity analysis

| Statistic | Examinees | Raters | Items |
|---|---|---|---|
| M (measure) | 0.37 | 0.14[a] | 0.00[b] |
| SD (measure) | 1.35 | 0.09 | 1.03 |
| M (SE) | 0.22 | 0.05 | 0.06 |
| Adj. (true) SD | 1.33 | 0.08 | 1.03 |
| Homogeneity index (Q) | 5,411.1** | 38.0** | 4,001.0** |
| *df* | 199 | 9 | 14 |
| Separation ratio (G) | 5.95 | 1.80 | 18.14 |
| Separation (strata) index (H) | 8.27 | 2.73 | 24.52 |
| Separation reliability (R) | .97 | .76 | > .99 |

**Note** [a] The rater facet was anchored by setting the expert rater's measure to 0.0 logits.
[b] The item facet was centered (i.e., this facet was constrained to have a mean element measure of 0.0 logits). ** $p < .01$.

**Table 2** Measurement results for the rater severity analysis

| Rater | Observed Average | Severity Measure | SE | Infit | Outfit |
|---|---|---|---|---|---|
| 1 | 0.52 | 0.26 | 0.05 | 1.00 | 0.99 |
| 5 | 0.53 | 0.23 | 0.05 | 1.02 | 1.03 |
| 3 | 0.54 | 0.19 | 0.05 | 1.00 | 1.00 |
| 9 | 0.54 | 0.19 | 0.05 | 0.98 | 0.96 |
| 7 | 0.54 | 0.18 | 0.05 | 1.01 | 1.03 |
| 2 | 0.54 | 0.14 | 0.05 | 1.02 | 1.05 |
| 4 | 0.55 | 0.12 | 0.05 | 1.01 | 1.01 |
| 6 | 0.56 | 0.07 | 0.05 | 1.00 | 1.05 |
| Expert | 0.57 | 0.00 | 0.05 | 0.98 | 0.97 |
| 8 | 0.57 | -0.02 | 0.05 | 0.97 | 0.98 |

**Note** Raters are ordered by severity measure (logits), from high (severe rater) to low (lenient rater).
Each rater scored the responses of 200 examinees to 15 dichotomous items.
The expert rater was anchored at 0.0 logits. Infit and outfit are mean-square fit statistics.

at a very high level (the *SE* estimates were all close to 0). Consequently, using a test statistic derived from the general Wald approach (Eckes, 2015, p. 61), all rater severity measures, except for the measures of Rater 6 and Rater 8, proved to differ significantly from the expert reference measure ($p < .01$).

The last two columns of Table 2 provide evidence on the extent to which the data fit the model used in the analysis. That is, the fit statistics indicate the extent to which the marks provided by a given rater matched the expected marks generated by the severity facets model defined in Equation 1. Specifically, rater infit is sensitive to unexpected ratings where the locations of a given rater and the elements of the other facets are closer together on the measurement scale; rater outfit is sensitive to unexpected ratings where the latent variable locations of that rater and the other elements are farther apart from each other (Eckes, 2015, 2019; Linacre, 2018b). In the present analysis, infit and outfit values were well within very narrow quality control limits (0.90/1.10), demonstrating a highly satisfactory fit between the data and the model (Linacre, 2002, 2018b).

***Rater-by-item interaction analysis***

The interaction model shown in Equation 2 was used to run a rater-by-item interaction analysis. Adopting an exploratory approach (Eckes, 2015), all 150 combinations of raters and items were scanned for significant differences between observed scores and expected scores, where the expected scores were derived from the main-effects severity facets model (Eq. 1). FACETS computed a bias statistic that was suited to examine the bias parameter estimate's statistical significance. Specifically, the bias statistic provided a test of the hypothesis that there was no item-related differential severity effect apart from measurement error. This statistic is approximately distributed as a $t$ statistic with $df = N - 1$ (where $N$ is the number of scores per rater).

Overall, the analysis revealed a very low level of differential severity: No more than four of the 150 bias terms proved to be statistically significant (i.e., $p < .05$). The raters involved were Rater 1 (concerning two items), Rater 4, and Rater 8. Using a Bonferroni adjustment procedure to control for the Type I error rate, three of these bias terms failed to reach the adjusted level of significance (i.e., $p < .0003$), leaving only Rater 4 as significantly biased (concerning a single item).

For illustrative purposes, Figure 2 displays the bias diagrams for these three raters. In each diagram, the bias statistic's values (in terms of $t$ values) are shown along the vertical axis, and the short-answer item numbers are shown along the horizontal axis. The expert's bias statistics are graphically included in the figure since this rater served as a reference.

For ease of interpretation, the bias diagrams shown in Figure 2 include the upper and lower quality control limits (dotted lines), conventionally drawn at $t$ values of –2 and +2, respectively. Generally, positive $t$ values mean that observed (total) scores were higher than expected based on the model, indicating a tendency toward leniency; negative $t$ values mean that observed (total) scores were lower than expected based on the model, indicating a tendency toward severity.

The expert rater provided marks that stayed well within the upper and lower control limits, thus exhibiting no item-related severity (or leniency) bias at all. Unlike the expert, Rater 1 showed a tendency toward severity when scoring examinee responses to Items 21 and 25. Table 3 presents the detailed results from the differential severity analysis for this rater and the other two raters.

Concerning Item 3, Rater 4 showed a highly significant difference between the observed score (92) and the expected score (120.29), indicating a strong differential severity effect when scoring examinee perfor-

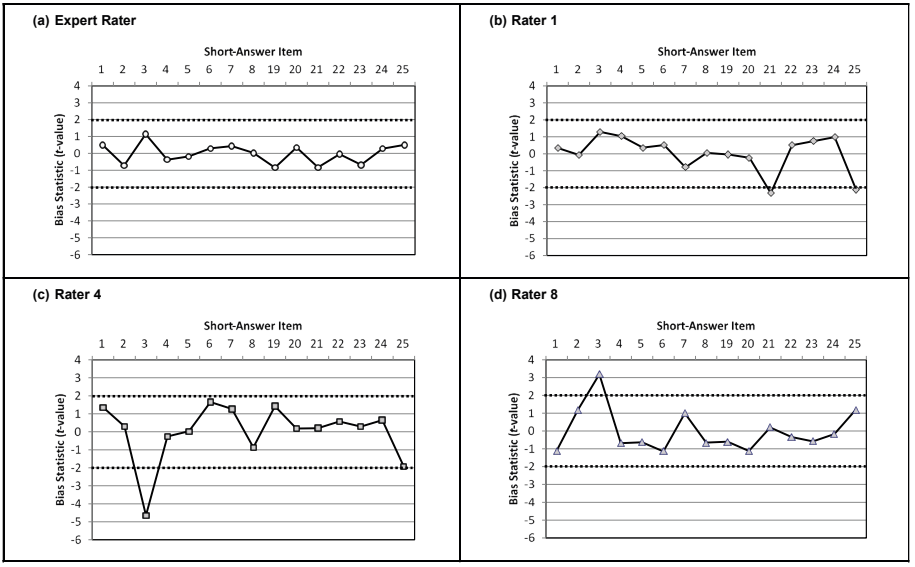


**Figure 2** Differential severity diagrams for the expert rater and three operational raters

**Table 3** Exemplary results from the rater-by-item differential severity analysis

| Rater | Item | N Scores | Observed Score | Expected Score | Bias Measure | SE | t | *p* |
|---|---|---|---|---|---|---|---|---|
| 1 | 21 | 198 | 68 | 81.79 | –0.40 | 0.17 | –2.29 | .023 |
| 1 | 25 | 197 | 76 | 88.59 | –0.35 | 0.17 | –2.07 | .040 |
| 4 | 3 | 196 | 92 | 120.29 | –0.78 | 0.17 | –4.70 | .000 |
| 8 | 3 | 196 | 144 | 125.27 | 0.58 | 0.18 | 3.16 | .002 |

**Note** The expected score was derived from the main-effects model (Eq. 1). The bias measure was estimated building on the interaction model (Eq. 2). Using the Bonferroni adjustment procedure, only the t statistic value for Rater 4 scorings on Item 3 proved to be significant (adjusted p value = .0003).

mance on this particular item; the resulting bias estimate was –0.78 logits ($SE$ = 0.17), with $t(195) = -4.70$, $p < .0001$. Considering the Rater 4's overall severity (0.12 logits; Table 2), this rater's severity as specifically related to Item 3 (the rater's local severity measure) was 0.90 logits (i.e., 0.12 logits – [–0.78 logits]). On the very same item, Rater 8 showed the opposite tendency, providing an observed score that was much greater (144) than expected (125.27), indicating a specific tendency toward leniency.

## Rater accuracy analysis

### *Wright map*

Figure 3 displays the accuracy-related measures for examinees, raters, and items in a common frame of reference. Unlike Figure 1, the second column ("+Examinees") presents the measures for examinees whose performances were easy to rate accurately at the top, and the measures for examinees whose performances were difficult to rate accurately at the bottom. The third column ("+Raters") compares the raters in terms of their ability to assign accurate scores to examinee performances. More accurate raters appear higher in the column, while less accurate raters appear lower. Finally, the fourth column ("+Items") arranges the items according to the easiness of scoring them accurately; short-answer items appearing higher in the column were easier to score accurately than those appearing lower.

```
+-----------------------------------------------+
|Measr|+Examinees  |+Raters     |+Items          |
|-----+------------+------------+----------------|
|     |Easy to mark|  Accurate  |  Easy to mark  |
|     | accurately |            |   accurately   |
|     |            |            |                |
|   4 + **         +            +                |
|     |            |            |                |
|     |            |            |                |
|     |            |            |                |
|     |            |            |                |
|     |            |            |                |
|     |            | 9          |                |
|     |            | 2  8       |                |
|   3 +            + 6          +                |
|     |            | 1  4  5  7 |                |
|     |            | 3          |                |
|     |            |            |                |
|     |            |            |                |
|     |            |            |                |
|     | **.        |            |                |
|   2 + *          +            +                |
|     |            |            |                |
|     |            |            | 4              |
|     | .          |            |                |
|     | *****      |            |                |
|     | *          |            |                |
|     |            |            | 23             |
|   1 + ****       +            +                |
|     | **         |            |                |
|     | *****.     |            |                |
|     | *.         |            | 5              |
|     | *****.     |            |                |
|     | *          |            |                |
|     | ******     |            | 22             |
|     | *******    |            | 2   8          |
*   0 * ******     *            *                *
|     | ******     |            | 19             |
|     | *****      |            | 20  24  25  6  |
|     | *****.     |            | 1              |
|     | ******.    |            | 7              |
|     | *******.   |            |                |
|     | *.         |            | 21             |
|     | *****      |            |                |
|  -1 + ****.      +            +                |
|     | **         |            |                |
|     | ***        |            | 3              |
|     | *          |            |                |
|     | *.         |            |                |
|     | .          |            |                |
|     |            |            |                |
|     | Difficult  |            |   Difficult    |
|  -2 + to mark a. + Inaccurate +  to mark accur.|
|-----+------------+------------+----------------|
|Measr| * = 2      |+Raters     |+Items          |
+-----------------------------------------------+
```

**Figure 3** Wright map for the rater accuracy analysis

The Wright map demonstrates that examinee performances varied widely in the easiness for operational raters to assign accurate scores. Excluding those four examinees whose performances were scored with perfect accuracy, the easiness measures showed a remarkable 3.73-logit spread.

Quite different from the accuracy-related examinee locations, the rater accuracy measures were far more homogeneous, clustering at the high-accuracy portion of the logit scale (remember that the rater facet was non-centered in this analysis). The overall high degree of accuracy manifested itself through an exact agreement rate between operational and expert scores of 92.8%; that is, only 1,844 scores out of a total of 25,758 scores assigned by operational raters differed from the expert's scores. The most accurate rater, that is, the rater showing the highest agreement with the expert, was Rater 9; the least accurate rater was Rater 3.

***Summary Rasch statistics***

The summary Rasch statistics shown in Table 4 lend substance to the conclusions drawn from the Wright map (Fig. 2). For each facet, the value of the homogeneity index *Q* was statistically significant, attesting to a pronounced heterogeneity among examinee, rater, and item measures, respectively. Moreover, the rater strata index revealed that there were almost three classes of raters that can be reliably distinguished in terms of their ability to score accurately. Finally, the rater separation reliability left no doubt that the operational raters, though all of them extensively-trained and well-experienced professionals, exhibited notable differences in their ability to provide accurate scores.

**Table 4** Summary Rasch statistics for the many-facet rater accuracy analysis

| Statistic | Examinees | Raters | Items |
|---|---|---|---|
| M (measure) | 0.00[a] | 2.99 | 0.00[a] |
| SD (measure) | 0.92 | 0.15 | 0.71 |
| M (SE) | 0.45 | 0.08 | 0.11 |
| Adj. (true) SD | 0.69 | 0.14 | 0.70 |
| Homogeneity index (Q) | 711.0** | 33.3** | 479.7** |
| *df* | 199 | 8 | 14 |
| Separation ratio (G) | 1.52 | 1.86 | 6.06 |
| Separation (strata) index (H) | 2.36 | 2.81 | 8.42 |
| Separation reliability (R) | .70 | .78 | .97 |

**Note** [a]The examinee and item facets were centered (i.e., these facets were each constrained to have a mean element measure of 0.0 logits). ** $p < .01$.

***Rater accuracy calibrations***

The results for the nine operational raters presented in Table 5 provide a detailed account of their accuracy measures and the fit of their marks to the Rasch model. There was a clear ordering from accurate to inaccurate raters and a very high agreement with model expectations as evidenced by the infit and outfit statistics. Compared to the other raters in the group, Rater 4 showed a somewhat heightened outfit value (1.14); conversely, Rater 1 tended to show a somewhat lessened outfit value (0.88), indicating a tendency to overfit the model.

***Rater-by-item interaction analysis***

The rater-by-item interaction analysis using the model shown in Equation 4 yielded a negligibly low level of differential accuracy: No more than three of the 135 bias terms proved to be statistically significant (i.e., $p < .05$). The raters involved were Rater 1, Rater 4, and Rater 6. Using a Bonferroni adjustment procedure to control for the Type I error rate, none of these bias terms reached the adjusted level of significance (i.e., $p < .0004$).

For illustrative purposes, Figure 4 displays the bias diagrams for these three raters. As before, each diagram includes the upper and lower quality control limits (dotted lines). In the present case, positive $t$ values indicate that a particular rater was more accurate than expected, given the rater's overall accuracy and the item's easiness to be scored accurately; negative $t$ values indicate a particular rater's item-specific tendency toward inaccuracy.

There was a commonality among the raters' item-specific differential accuracy: The deviations from model expectations referred to the same item, that is, Item 25, albeit in opposite directions. Thus, Rater 1 and Rater 4 scored this particular item less accurately than expected; Rater 6 scored the item more accurately than expected.

**Table 5** Measurement results for the rater accuracy analysis

| Rater | Observed Average | Accuracy Measure | SE | Infit | Outfit |
|---|---|---|---|---|---|
| 9 | 0.95 | 3.28 | 0.08 | 0.99 | 0.96 |
| 8 | 0.94 | 3.15 | 0.08 | 1.00 | 0.89 |
| 2 | 0.94 | 3.11 | 0.08 | 0.99 | 0.96 |
| 6 | 0.93 | 2.98 | 0.07 | 1.00 | 1.01 |
| 7 | 0.93 | 2.93 | 0.07 | 1.00 | 1.08 |
| 4 | 0.92 | 2.90 | 0.07 | 1.03 | 1.14 |
| 5 | 0.92 | 2.89 | 0.07 | 1.00 | 1.08 |
| 1 | 0.92 | 2.86 | 0.07 | 1.00 | 0.88 |
| 3 | 0.92 | 2.79 | 0.07 | 1.00 | 0.96 |

**Note** Raters are ordered by accuracy measure (logits), from high (accurate rater) to low (inaccurate rater). Infit and outfit are mean-square fit statistics.

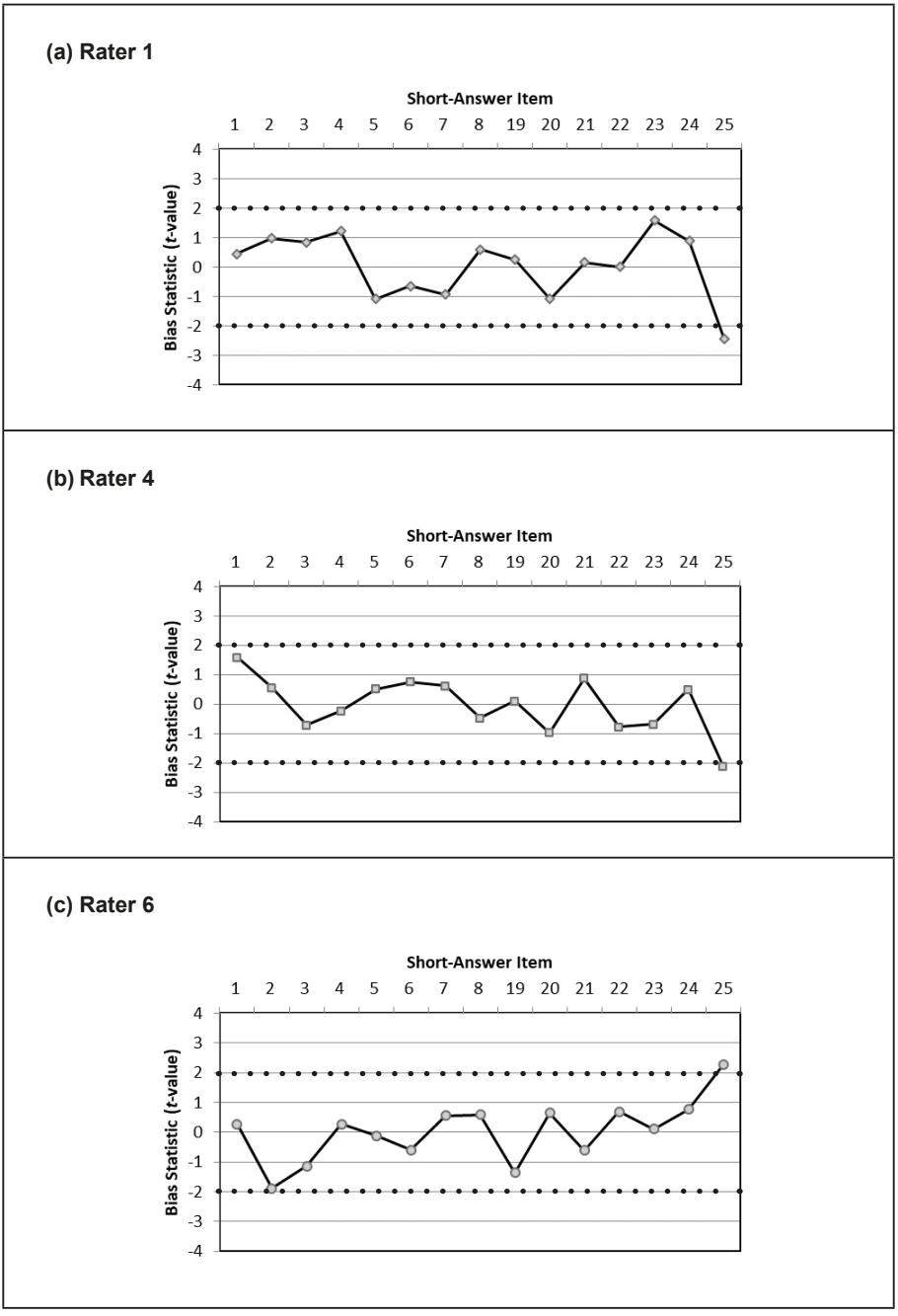


**Figure 4** Differential accuracy diagrams for three operational raters

Table 6 presents the detailed results from the differential accuracy analysis for these three raters.

For example, Rater 1 showed a significant difference between the observed score (167) and the expected score (176.58), demonstrating less accuracy than expected when assigning scores to examinee performances on Item 25; the resulting bias estimate was −0.52 logits ($SE$ = 0.21), with $t(193) = -2.43$, $p$ = .016. Based on this particular rater's overall accuracy (2.86 logits; Table 4), his or her local measure was 2.34 logits (i.e., 2.86 logits + [−0.52 logits]); that is, Rater 1 was specifically less accurate when scoring responses to Item 25. On the very same item, Rater 6's local measure was 3.86 logits (i.e., 2.98 logits + [0.88 logits]), indicating a tendency toward still greater accuracy when scoring performances on this particular item (remember that the rater facet in the rater accuracy analysis had a positive orientation).

## Summary and Discussion

The main purpose of the present research was to investigate the psychometric quality of scores assigned to examinee performances on the listening section of a large-scale, high-stakes assessment instrument (i.e., TestDaF). Two rating quality dimensions were put under scrutiny building on a facets modeling approach: rater severity and rater accuracy. The research questions (RQs) and the answers provided by this study can be summarized as follows.

RQ1 addressed the extent to which raters showed differences in their severity or leniency when scoring examinee responses. Building on a many-facet Rasch measurement model, the analysis revealed a statistically significant amount of severity variation among raters. Though relatively small, the differences in rater severity measures were not negligible. In particular, the rater separation (strata) index and the rater separation reliability suggested that the raters in the present sample were not interchangeable.

Concerning RQ2, the rater-by-item-interaction analysis was to provide insight into the extent to which raters showed varying levels of severity or leniency across

**Table 6** Exemplary results from the rater-by-item differential accuracy analysis

| Rater | Item | N Scores | Observed Score | Expected Score | Bias Measure | SE | *t* | *p* |
|---|---|---|---|---|---|---|---|---|
| 1 | 25 | 194 | 167 | 176.58 | −0.52 | 0.21 | −2.43 | .016 |
| 4 | 25 | 194 | 169 | 177.14 | −0.46 | 0.22 | −2.10 | .037 |
| 6 | 25 | 194 | 187 | 178.30 | 0.88 | 0.39 | 2.26 | .025 |

**Note** The expected score was derived from the main-effects model (Eq. 3). The bias measure was estimated building on the interaction model (Eq. 4). Using the Bonferroni adjustment procedure, none of the t-statistic values proved to be significant (adjusted p value = .0004).

short-answer questions. As it turned out, the level of differential rater severity was very low overall. Only one operational rater proved to be significantly biased toward just one of the 15 items considered in the analysis. The expert rater provided scores that were very well in line with model expectations, demonstrating a uniform level of severity or leniency across items (like the great majority of operational raters).

RQ3 shifted the focus on the accuracy of the scores provided by the operational raters. There were two parts to this question: (a) How well did the operational scores agree with the scores provided by the expert rater? (b) How much did the operational raters differ from each other in their accuracy measures? The answer to the first part was clear: There was an overall high degree of scoring accuracy; that is, the scores provided by operational raters agreed with those provided by the expert in the vast majority of cases (the agreement rate was nearly 93%). Regarding the second part, it became evident that the level of scoring accuracy was far from perfect: there were statistically significant differences between raters in their ability to provide accurate scores. Thus, the rater separation index and the rater separation reliability confirmed that some raters were significantly more accurate than others.

Finally, the focus of RQ4 was on the extent to which raters showed evidence of differential functioning in terms of varying levels of accuracy across short-answer questions. The answer was straightforward: There was no evidence for item-related differential rater accuracy in the present sample. If at all, there was a slight tendency for just one item to be scored less accurately than expected by two raters and to be scored more accurately than expected by one rater.

These findings have at least three implications. First, whenever feasible, a many-facet Rasch analysis should be run to examine the psychometric quality of the scores that the raters provided. In the case of significant between-rater severity variation, the final assessment outcomes could then be based on adjusted (fair) scores to compensate for rater severity differences.[3] Such a score adjustment procedure presupposes that the scoring design ensures connectedness of the data; that is, all elements of all facets involved need to be linked to each other directly or indirectly such that they can be represented in a common frame of reference (Eckes, 2015; Engelhard & Wind, 2018). The present study utilized a fully crossed design, where all raters scored listening performances on all short-answer questions, resulting in a maximally connected data set. A second precondition for score adjustments is that the raters providing scores on the listening section demonstrate a sufficient fit to Rasch model expectations. Rater fit statistics like infit and outfit mean-square indices can be used to check this requirement (Eckes, 2015, 2019; Engelhard & Wind, 2018). For the present data, rater fit statistics provided clear evidence of good data–model fit. Hence, both requirements were met.

Second, even extensive rater training, as routinely implemented by the TestDaF development team before the operational scoring sessions start, is not sufficient to make the raters function interchangeably. This finding is reminiscent of the conclusions drawn from many studies evaluating the effectiveness of rater training concern-

3 In small-scale or classroom contexts, the proposed procedure will be challenging to implement. However, in large-scale and notably high-stakes situations, score adjustment based on facets models may help increase the fairness of listening assessments that use clerical marking.

ing writing and speaking assessments (Eckes, 2015, 2019). In other words, even in the context of clerical marking of listening performances, raters usually do not behave like "scoring machines", to borrow a term from Linacre (2018b). Nonetheless, rater training and monitoring activities can be enhanced in at least two ways using individualized feedback gained from a facets analysis. Rater severity statistics may help raters reconsider their understanding of the targeted listening ability and their performance expectations. Rater accuracy statistics may help raters check adherence to the specifications contained in the scoring guide and, if necessary, recalibrate their own, taken for granted scoring standards. Of course, making use of information on scoring accuracy in this way requires an expert or group of experts to provide consensus scores on examinee responses to short-answer questions. However, the proposed improvements to training procedures may be much more readily implemented in large-scale listening assessments than in smaller-scale contexts.

Third, when particular examinees' responses or particular short-answer questions are more difficult to score accurately than others, it would be helpful to know why these accuracy differences occurred in the first place. A wide variety of factors potentially impacting on the difficulty of listening items have been discussed in the literature (Brunfaut, 2016; Green, 2017). Unfortunately, knowledge about exactly which factors have an adverse or beneficial impact on scoring accuracy is seriously lacking. Rater accuracy facets models appear well-suited to contribute to filling this gap by pointing to individual facets or combinations of facets that are closely associated with the occurrence of scoring problems.

Finally, a word of caution should be added here. Given that the size of the examinee sample was relatively small, any of the findings and their implications discussed above need to be corroborated using a much larger sample from live language assessments. Nonetheless, efforts were made to ensure that the present sample of listening performances closely mirrored the critical decision-making regions of the latent proficiency scale. Also, the rater accuracy analysis was based on a comparison of the scores provided by a small group of operational raters to the scores of just a single expert rater. Using a greater number of operational raters and a group of experts would undoubtedly increase the reliability of the findings presented here. It should be noted, however, that the basic approach adopted in this research may well serve as a paradigm for rater accuracy studies in the field of listening assessment on a larger scale.

## Conclusion

Buck (2018) noted that listening "is still the most neglected of the traditional four skills. This is unfortunate because, in a number of ways, listening can be regarded as the most fundamental skill of all" (p. xi). In a similar vein, Harding et al. (2015) characterized listening as "one of the most under-researched aspects of assessment, reflecting its 'Cinderella' status among the four skills" (p. 326). The present study recognized the critical role of this skill by (a) conceptualizing the assessment of listening through scoring of examinee responses to short-answer questions as an instance of the more general class of rater-mediated assessments, much like scoring writing or speaking performances, and (b) adopting a

facets modeling approach to the analysis of rating quality with a focus on rater severity and rater accuracy. There was evidence that in a standardized listening assessment using short-answer questions along with a detailed scoring guide, raters exhibited differences in the severity and accuracy of the scores assigned to examinee responses. Though at a relatively low level overall, these differences were far from negligible. The present findings highlight the need to implement procedures for analyzing and evaluating rater-mediated assessments of listening ability on a routine basis, particularly in high-stakes contexts. There is also a demand for research into the factors that contribute to rater variability in listening assessments.

## References


Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, *19*(4), 369–394. https://doi.org/10.1191/0265532202lt236oa

Brunfaut, T. (2016). Assessing listening. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 97–112). de Gruyter. https://doi.org/10.1515/9781614513827-009

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

Buck, G. (2018). Preface. In G. J. Ockey & E. Wagner, *Assessing L2 listening: Moving towards authenticity* (pp. xi–xvi). John Benjamins. https://doi.org/10.1075/lllt.50

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford University Press.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*, *31*, 39–61. https://doi.org/10.1177/0265532213492969

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Peter Lang. https://doi.org/10.3726/9783653048445

Eckes, T. (2017). Rater effects: Advances in item response modeling of human ratings – Part I (Guest Editorial). *Psychological Test and Assessment Modeling*, *59*(4), 443–452. https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2017_20171218/03_Eckes.pdf

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment: Vol. 1. Fundamental techniques* (pp. 153–175). New York, NY: Routledge. https://doi.org/10.4324/9781315187815

Eckes, T., & Althaus, H.-J. (2020). Language proficiency assessments in higher education admissions. In M. E. Oliveri & C. Wendler (Eds.), *Higher education admission practices: An international perspective* (pp. 256–275). Cambridge University Press. https://doi.org/10.1017/9781108559607

Elder, C. (2017). Language assessment in higher education. In E. Shohamy, I. G. Or & S. May (Eds.), *Language testing and assessment* (3rd ed., pp. 271–286). Springer.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*(1), 56–70. https://doi.org/10.1111/j.1745-3984.1996.tb00479.x

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). Erlbaum.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Engelhard, G., Wang, J., & Wind, S. A. (2018). A tale of two models: Psychometric and cognitive perspectives on rater-mediated assessments using accuracy ratings. *Psychological Test and Assessment Modeling*, *60*(1), 33–52. https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2018_20180323/3_PTAM_Engelhard__Wang___Wind__2018-03-10__1855.pdf

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge. https://doi.org/10.4324/9781315766829

Fan, J., & Knoch, U. (2019). Fairness in language assessment: What can the Rasch model offer? *Papers in Language Testing and Assessment*, *8*(2), 117–142. http://www.altaanz.org/uploads/5/9/0/8/5908292/8_2_s5_fan_and_knoch.pdf

Field, J. (2019). *Rethinking the second language listening test: From theory to practice*. Equinox.

Geranpayeh, A. (2013). Scoring validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening: Research and practice in assessing second language listening* (pp. 242–272). Cambridge University Press.

Green, R. (2017). *Designing listening tests: A practical approach*. Palgrave Macmillan. https://doi.org/10.1057/9781349687718

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3), 317–336. https://doi.org/10.1177/0265532214564505

In´nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26*(2), 219–244. https://doi.org/10.1177/0265532208101006

Jin, K.-Y., & Wang, W.-C. (2018). A new facets model for rater's centrality/extremity response style. *Journal of Educational Measurement*, *55*(4), 543–563. https://doi.org/10.1111/jedm.12191

Kang, T., Gutierrez Arvizu, M. N., Chaipuapae, P., & Lesnov, R. O. (2019). Reviews of academic English listening tests for non-native speakers. *International Journal of Listening*, *33*(1), 1–38. https://doi.org/10.1080/10904018.2016.1185210

Kunnan, A. J. (2012). Language assessment for immigration and citizenship. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 162–177). Routledge.

Lane, S. (2019). Modeling rater response processes in evaluating score meaning. *Journal of Educational Measurement*, *56*(3), 653–663. https://doi.org/10.1111/jedm.12229

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878. https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2018a). Facets Rasch measurement computer program (Version 3.81) [Computer software]. Winsteps.com.

Linacre, J. M. (2018b). *A user's guide to FACETS: Rasch-model computer programs*. Winsteps.com. http://www.winsteps.com/facets.htm

McNamara, T. F. (2000). *Language testing*. Oxford University Press.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment: The role of measurement*. Oxford University Press.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422. http://jampress.org/pubs.htm

Norris, J., & Drackert, A. (2018). Test review: TestDaF. *Language Testing*, *35*(1), 149–157. https://doi.org/10.1177/0265532217715848

Ockey, G. J., & Wagner, E. (2018). *Assessing L2 listening: Moving towards authenticity*. John Benjamins. https://doi.org/10.1075/lllt.50

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press. (Original work published 1960)

Robitzsch, A., & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods and implementation in R. *Psychological Test and Assessment Modeling*, *60*(1), 101–138. https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2018_20180323/6_PTAM_IRMHR_Main__2018-03-13_1416.pdf

Rost, M. (2016). *Teaching and researching listening* (3rd ed.). Routledge.

Song, M.-Y. (2011). Note-taking quality and performance on an L2 academic listening test. *Language Testing*, *29*(1), 67–89. https://doi.org/10.1177/0265532211415379

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185–201. https://doi.org/ 10.1111/j.1745-3984.1987.tb00274.x

Wind, S. A., & Engelhard, G. (2012). Examining rating quality in writing assessment: Rater agreement, error, and accuracy. *Journal of Applied Measurement*, *13*(4), 321–335. http://jampress.org/pubs.htm

Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, *18*(4), 278–299. https://doi.org/10.1016/j.asw.2013.09.002

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161–192. https://doi.org/10.1177/0265532216686999

Wolfe, E. W., Jiao, H., & Song, T. (2015). A family of rater accuracy models. *Journal of Applied Measurement*, *16*(2), 153–160. http://jampress.org/pubs.htm

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31–37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

***Corresponding author:***
***Thomas Eckes, PhD***
*TestDaF Institute*
*University of Bochum*
*Universitätsstr. 134*
*44799 Bochum,*
*Germany*
*thomas.eckes@gast.de*