THE TESTDAF IMPLEMENTATION OF THE SOPI:

DESIGN, ANALYSIS, AND EVALUATION OF A SEMI-DIRECT SPEAKING TEST

Thomas Eckes, TestDaF Institute, Germany

Abstract

The Test of German as a Foreign Language (*Test Deutsch als Fremdsprache*, TestDaF) is a standardized test designed for foreign learners of German who plan to study in Germany or who require recognized certification of their language ability. In its speaking section, the TestDaF makes use of an adapted version of the Simulated Oral Proficiency Interview (SOPI; Kenyon, 2000; Stansfield & Kenyon, 1988). The present paper discusses the design of the TestDaF speaking instrument and illustrates the measurement approach adopted in order to analyze and evaluate this particular implementation of the SOPI. General requirements for the operational use of the speaking test are delineated and the specific test format is described. The main part deals with the analysis and evaluation of speaking performance ratings obtained from a live TestDaF examination. The paper concludes with perspectives for future test development.

The TestDaF: Purpose and Scope

The TestDaF allows foreign students applying for entry to an institution of higher education in Germany to prove their knowledge of German while still in their home country. Test tasks and items are continuously developed, analyzed, and evaluated at the TestDaF Institute (Hagen, Germany); examinee performance is also centrally scored at this institute (Eckes, 2008a; Eckes et al., 2005; Grotjahn, 2004; see also www.testdaf.de).

The TestDaF measures the four language skills (i.e., reading, listening, writing, and speaking) in separate sections. Examinee performance in each section is related to one of three levels of language ability in the form of band descriptions; these levels (*TestDaF-Niveaustufen*, TestDaF levels, or TDNs for short) are TDN 3, TDN 4, and TDN 5. The TDNs 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 ability at an intermediate to high level (see Kecker & Eckes, in press). There is no differentiation among lower ability levels; it is just noted that TDN 3 has not yet been achieved (below TDN 3).

The TestDaF is officially recognized as a language entry exam for students from abroad. Examinees who have achieved at least TDN 4 in each section are eligible for admission to a German institution of higher education (see Eckes et al., 2005).

In April 2001, the TestDaF was administered worldwide for the first time. Until the end of 2009, more than 106,000 students had taken this test. The number of test administrations per year increased from two exams in 2001 to nine exams until present (including three separate exams in the People's Republic of China). Table 1 portrays the growth of the TestDaF candidature from 2001 to 2009, as well as the number of test centers and test countries during this period (see also TestDaF Institute, 2010).

Table 1- *Growth of TestDaF Candidature, Test Centers, and Test Countries*

Year	Test Takers	Test Centers	Test Countries
2001	1,190	81	34
2002	3,582	154	48
2003	7,498	211	65
2004	8,982	261	70
2005	11,052	275	72
2006	13,554	309	74
2007	15,389	318	75
2008	16,882	330	78
2009	18,059	329	77

Speaking Test Requirements

The construct underlying the TestDaF draws on Bachman and Palmer's (1996) model of communicative language ability. More specifically, for each language skill the following areas of language knowledge are taken into account: grammatical knowledge (not assessed separately, but indirectly addressed in each of the TestDaF sections), textual knowledge (coherence/cohesion, rhetorical functions, conversational organization), functional knowledge (ideational, manipulative, heuristic functions), and sociolinguistic knowledge (registers, idiomatic expressions, cultural references). In addition, TestDaF tasks are intended to tap into areas of strategic competence (goal setting, assessment, planning).

Considering the test construct, the TestDaF speaking section is designed as a performance-based instrument, assessing the examinees' ability to communicate appropriately in typical situations of university life. Accordingly, item writers are instructed to produce tasks that elicit language use

relevant to, and characteristic of, this specific context. Beyond this basic, construct-driven requirement for test development, the speaking test has to meet a number of more practical demands. To begin with, trained examiners/interviewers of German as a foreign language are not readily available in many regions of the world. Therefore, the speaking test is to be administered without the use of on-site examiners. Moreover, for reasons of test security as well as cost-effectiveness in terms of test delivery and administration, a large number of examinees are to be tested worldwide on a single test version at the same day.

Further demands relate to issues of standardization, reliability, and validity of the speaking test. Thus, the requirement of standardization says that each examinee receives the same instructions, prompts, and questions as any other examinee taking the test. Ideally, there should be no variation in examiner and/or interviewer input. A high degree of standardization is a prerequisite condition for assuring high reliability of the assessment outcomes. High assessment reliability implies that variation in assessment outcomes that is due to random measurement error is negligibly small. High reliability, in turn, is a necessary but not sufficient condition for high validity. For example, when speaking performance is scored by human raters, differences in rater severity or leniency are generally observed, resulting in variation in assessment outcomes that is not associated with the performance of examinees. As a consequence hereof, examinee speaking ability is not adequately assessed, lowering the validity of the conclusions drawn from the assessment outcomes. It is obvious that in cases like this the assessment instrument has limited fairness as well; that is, some examinees may benefit from being rated by a lenient rater, whereas others may suffer from bad luck in terms of getting a severe rater.

Speaking Test Design

Most of the requirements outlined above are met by the SOPI format. The SOPI is a type of semi-direct speaking test (Luoma, 2004; Qian, 2009), developed in the 1980s at the Center of Applied Linguistics in Washington, DC (for reviews, see Kenyon, 2000; Kuo & Jiang, 1997). This testing format was designed to model the nature of the Oral Proficiency Interview (OPI) used by the American Council on the Teaching of Foreign Languages (ACTFL). Whereas the OPI is a face-to-face interview, the SOPI relies on pre-recorded prompts and a printed test booklet to elicit language from the examinee.

Early research suggested that the SOPI is a reliable and valid technology-based alternative to the OPI (Kenyon, 2000; Stansfield & Kenyon, 1992). For example, Stansfield and Kenyon (1992) performed a number of correlation studies and concluded that “the OPI and the SOPI are close enough in the way they measure general speaking proficiency that they may be viewed as parallel tests delivered in two different formats” (p. 359). However, researchers have also provided evidence that direct and semi-

direct speaking tests may tap different language abilities, in particular, interactive versus monologic speaking ability (see, e.g., O'Loughlin, 2001; Shohamy, 1994; see also Galaczi, this volume). Qian (2009) added another facet to the comparison between direct and semi-direct speaking tests. The author studied affective effects of direct and semi-direct modes for speaking assessment on test-takers and found that "a large proportion of the respondents in the study were quite willing to accept both testing modes" (p. 123).

The typical full-length SOPI, used to assess examinees from the Novice to the Superior level on the ACTFL scale, comprises a total of 15 tasks measuring general speaking ability in a foreign or second language. The SOPI currently in use with the TestDaF is an adapted version consisting of seven speaking tasks (including a warm-up task) tailored to the German academic context. Following this format, the speaking test is administered via audio-recording equipment using prerecorded prompts and printed test booklets. That is, during testing the examinee listens to directions for speaking tasks from a master tape or CD while following along in a test booklet; as the examinee responds to each task, his or her speaking performance is recorded on a separate response tape or CD. Testing time is about 30 minutes. Before being put to operational use, each speaking task is carefully examined in an elaborate evaluation process comprising piloting and trialling stages (see Eckes, 2008a). Table 2 provides an overview of the main features characterizing the TestDaF speaking section.

Table 2 - *Overview of the TestDaF Speaking Assessment Instrument*

Feature	Description
Construct	Ability to communicate appropriately in typical situations of university life
Format	Semi-direct (SOPI); adapted version (German academic context); seven tasks (one warm-up, two tasks each for TDN 3, TDN 4, and TDN 5); testing time is 30 min.
Administration	Tape-mediated or computer-assisted; prompts are pre-recorded and text based (printed test booklet); the examinee speaks into a microphone while the response is recorded
Roles/Situations	Examinees act out themselves in conversations with other students, employees at university, professors, etc.; the conversations are situated in seminars, language courses, cafeteria, etc.
Register	formal, informal, semi-formal
Rating	Experienced and trained raters; analytic rating scheme; top-down rating procedure (performance on top-level TDN 5 tasks rated first)

Speaking tasks are presented to all examinees in the same, fixed order. In the first task, the "warm-up", the examinee is asked to make a simple request; performance on this task is not rated. The other tasks focus on situation-related communication (e.g., obtaining and supplying information), relate to "describing", or deal with "presenting arguments". Two tasks each probe into one of the relevant proficiency levels (i.e., TDN 3, TDN 4, or TDN 5). The test ends with a "wind-down" consisting of a less-challenging task intended to put the examinees at ease before leaving the examination. Table 3 shows how the speaking tasks progress in terms of the challenge (i.e., TDN level) they pose to examinees.

Table 3 - *Progression of Speaking Task Difficulty Level*

	Task No.						
Level	1	2	3	4	5	6	7
TDN 3	x	x					x
TDN 4			x		x		
TDN 5				x		x	

Note. Each "x" means that a given task is presented at the level indicated by the row. Task No. 1 is a warm-up task; performance on this task is not rated.

In each task, examinees are asked to act out themselves in simulated conversations with other students, employees at a university office, lecturers, professors, and so on. These conversations are typically situated in familiar academic settings, including seminars, language courses, and cafeterias. According to the different contexts of language use the relevant registers cover a wide range of formal, informal, and semi-formal varieties. Table 4 shows in a summary fashion which kind of speech act each task requires from the examinees. In the instructions to the test, examinees are explicitly asked to take the specific content of each task into account when responding. Figure 1a and Figure 1b present examples of speaking tasks aiming at proficiency levels TDN 3, TDN 4, and TDN 5, respectively. These tasks are taken from a TestDaF sample speaking test that is available online for test preparation purposes at www.testdaf.de (see the links "Für Teilnehmerinnen und Teilnehmer", "Vorbereitung", "Modellsatz 02").

Table 4 - *Speaking Tasks, TDNs, and Required Speech Acts*

Task No.	TDN	Speech Act
1	3	Asking for information
2	3	Reporting / describing cultural facts
3	4	Describing a diagram or chart
4	5	Commenting on a topic / balancing pros and cons (socio-political/sociocultural area)
5	4	Stating one‘s opinion on a particular topic (personal area)
6	5	Forming / presenting hypotheses based on a diagram
7	3	Giving advice / giving reasons

Note. Task No. 1 is a warm-up task; performance on this task is not rated.

Figure 1a. Sample speaking test. Task 2 (at TDN 3) is shown in the upper half, Task 5 (at TDN 4) in the lower half of page 69.

Figure 1b. Sample speaking test. Task 6 (at TDN 5) is shown in the upper half, the chart that this task refers to in the lower half of page 70.

Ihr Studienfreund Martin möchte aus der Wohnung seiner Eltern ausziehen und sucht deshalb eine neue Wohnung. Er fragt Sie, wie lange die jungen Leute in Ihrem Heimatland bei ihren Eltern leben.

Beschreiben Sie,
– wann junge Menschen in Ihrem Heimatland von zu Hause ausziehen und
– warum sie ihr Elternhaus verlassen.

Sie: Vorbereitungszeit — **1 Minute**

Martin: ...

Sie: Sprechzeit — **1 Minute**

Ihr Freund Steffen muss während seines Studiums ein Praktikum machen. Er hat zwei Möglichkeiten: Steffen kann das Praktikum entweder in der Firma seiner Eltern absolvieren. Oder er macht sein Praktikum in einem anderen Betrieb. Steffen fragt Sie nach Ihrer Meinung.

Sagen Sie Steffen, wozu Sie ihm raten:
– Wägen Sie Vorteile und Nachteile der beiden Möglichkeiten ab.
– Begründen Sie Ihre Meinung.

Sie: Vorbereitungszeit — **2 Minuten**

Steffen: ...

Sie: Sprechzeit — **1 Minute 30 Sekunden**

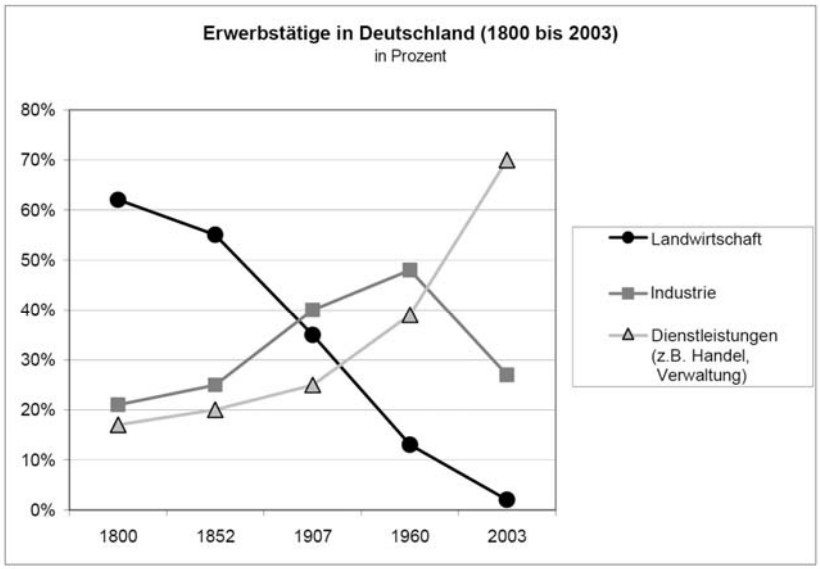
In Ihrem Wirtschaftsseminar geht es heute um die Veränderungen im Bereich Erwerbstätigkeit in Deutschland. Ihre Dozentin, Frau Dr. Maier, hat eine Grafik verteilt, die zeigt, in welchen Wirtschaftsbereichen die Menschen arbeiten. Frau Dr. Maier bittet Sie, Ihre Überlegungen zu Gründen der bisherigen Entwicklung und zur zukünftigen Entwicklung vorzutragen.

Nennen Sie mögliche Gründe für die dargestellte Entwicklung.
Stellen Sie dar, welche Entwicklung Sie für die Zukunft erwarten.
Begründen Sie Ihre Überlegungen anhand der Grafik.

Sie: Vorbereitungszeit — **3 Minuten**

Frau Dr. Maier: ...

Sie: Sprechzeit — **2 Minuten**



Nach: Statistisches Bundesamt, Wiesbaden 2004

All speaking tasks follow the same general design. Thus, each task has two clearly separated parts. In the first part, the communicative situation is described, and the examinee is told what to do. The examinee can read the task along in the test booklet. The booklet also shows the time to plan a response. Depending on the task, planning time ranges from 30s to 180s. During this time the examinee is allowed to take notes. In the second part, the "interlocutor" speaks, and the examinee is asked to listen and to respond after that. The time to respond is shown in the booklet. Again depending on the task, response time ranges from 30s to 120s. The examinee is free to stop responding before the response time is over.

Rating Speaking Test Performance

In direct speaking tests like the OPI, interviewer behavior is likely to exert considerable influence on the assessment outcomes, thus contributing to construct-irrelevant variance in examinee scores (see, e.g., Bachman, 2002; Brown, 2005; O'Sullivan, 2008). Due to its format, the SOPI eliminates any examiner or interviewer variability. Another important construct-irrelevant source of variability, however, remains largely unaffected by the SOPI format. This source relates to characteristics of the raters evaluating the quality of examinee responses after the test is completed.

Raters involved in rating examinee performance on TestDaF speaking tasks are subjected to elaborate training and monitoring procedures. Yet, extensive rater training sessions and detailed rater monitoring and feedback practices notwithstanding, rater variability generally remains a major unwanted component of examinee score variance (see, e.g., Hoyt & Kerns, 1999; Knoch, Read & von Randow, 2007; Lumley & McNamara, 1995; O'Sullivan & Rignall, 2007; Weigle, 1998). One reason why it is so difficult, if not impossible, to reduce rater variability to an acceptably low level in most assessment contexts is that this variability can manifest itself in a number of ways. For example, rater variability may take the form of (a) differences in rater severity or leniency, (b) differences in raters' understanding and use of rating scale categories, (c) differences in the kind of performance features raters attend to, (d) differences in the way raters interpret and use scoring criteria, or (e) various interactions of raters with examinees, tasks, criteria, and other facets of the assessment situation (see, e.g., Brown, 2005; Eckes, 2008b, 2009b; McNamara, 1996; Lumley, 2005).

The usual, or standard, approach to resolving the rater variability problem, especially in high-stakes assessments, consists of three components: rater training, independent ratings of the same performance by two or more raters (repeated ratings), and establishing interrater reliability. Since this approach has been shown to encounter difficulties of the kind mentioned above, another approach that has attracted a great deal of attention lately is to eliminate rater variability altogether by replacing human rating with

fully automated scoring (see, e.g., Chapelle & Chung, 2010; Shermis & Burstein, 2003; Williamson, Bejar & Mislevy, 2006).

For assessing writing performance, automated scoring systems have been in place for quite a while. Recent examples include *e-rater*, a system operationally used with the TOEFL iBT writing section (Attali, 2007; Weigle, 2010), the *Intelligent Essay Assessor* (IEA), used with the Pearson Test of English (PTE) Academic (Pearson, 2009), or *IntelliMetric* (Elliot, 2003; Wang & Brown, 2007). With some delay, automated scoring systems for assessing speaking performance have followed suit, such as *SpeechRater* (Xi, Higgins, Zechner & Williamson, 2008; Zechner, Higgins & Xi & Williamson, 2009), or the PTE Academic speaking test, which makes use of Ordinate technology (Bernstein & Cheng, 2007; see also van Moere, this volume).

The approach adopted within the context of the TestDaF speaking test is to continue employing human raters, but to compensate for differences in rater severity by means of a measurement approach that is based on the many-facet Rasch model (Linacre, 1989; Linacre & Wright, 2002). This approach is discussed in some detail next.

Speaking Test Analysis and Evaluation

The many-facet Rasch measurement (MFRM) model allows the researcher to examine more variables (or "facets") than the two that are typically included in a paper-and-pencil testing situation (i.e., examinees and items). Thus, in speaking performance assessments, additional facets that may be of particular interest refer to raters, speaking tasks, and scoring criteria. Within each facet, each element (i.e., each individual examinee, rater, task, or criterion) is represented by a parameter. These parameters denote distinct attributes of the facets involved, such as proficiency (for examinees), severity (for raters), and difficulty (for tasks or criteria).

Viewed from a measurement perspective, an appropriate approach to the analysis of many-facet data would involve three general steps. Step 1 refers to a careful inspection of the assessment design; that is, relevant issues to be considered at this stage concern the sample of examinees at which the assessment is targeted, the selection of raters to be used in the assessment, the nature of the speaking tasks, and many others. Step 2 concerns the specification of an appropriate measurement model; for example, determining the facets to be examined, or defining the structure of the rating scale. Step 3 calls for implementing that model in order to provide a fine-grained analysis and evaluation of the functioning of each of the facets under consideration (for a detailed discussion, see Eckes, 2009a, in press).

In what follows, I illustrate relevant features of the many-facet Rasch analysis routinely applied to the TestDaF rater-mediated system of speaking performance assessment (see also Eckes, 2005). The

database consisted of ratings of examinee performance on a speaking test as part of a live exam that took place in July 2008.

Examinees

The speaking test was administered to 1,771 participants (1,072 females, 699 males). Participants' mean age was 24.62 years (*SD* = 5.02); 87.0% of participants were aged between 18 and 30 years.

There were 152 TestDaF test centers involved in this administration (104 centers in Germany, 48 centers in 33 foreign countries). In terms of the number of examinees, the following five national groups ranked highest (percentage in parentheses): People's Republic of China (9.8%), Russia (9.4%), Ukraine (6.3%), Turkey (6.1%), Poland (4.6%).

Raters

Thirty-seven raters participated in the scoring of examinee speaking performance. Raters were all experienced teachers and specialists in the field of German as a foreign language, and were systematically trained and monitored to comply with scoring guidelines.

Procedure

Ratings of examinee speaking performance were carried out according to a detailed catalogue of performance aspects comprising eight criteria. The first two criteria (*comprehensibility, content*) were more holistic in nature, referring to the *overall impression* upon first listening to the oral performance, whereas the others were more of an analytic kind, referring to various aspects of *linguistic realization* (*vocabulary*, *correctness*, *adequacy*) and *task fulfillment* (*completeness*, *argumentation*, *standpoint*). On each criterion, examinee performance was scored using the four-point TDN scale (with categories *below TDN 3*, *TDN 3*, *TDN 4*, *TDN 5*).

Ratings were provided according to a top-down procedure. Thus, raters started with rating examinee performance on the two most challenging tasks (i.e., TDN 5 tasks). If the examinee was clearly a top-level speaker, then it was not necessary for the rater to listen to the examinee's performances on any of the lower-level tasks. Otherwise, the performance ratings were continued at TDN 4. If the examinee was clearly a speaker at that level, then the rating was finished; if not, examinee performances on the least-challenging tasks were also rated. In general, this procedure serves to increase rater efficiency and to prevent rater fatigue or waning of rater attentiveness. It is particularly efficient when used with digitally recorded speaking performances saved on CD, allowing the raters to skip forward and back within the audio file to quickly locate the next performance to be rated.

Each examinee performance was rated by a single rater. Such a rating design calls for measures to satisfy the precondition of connectivity of the resulting sparse data matrix. That is, all raters, examinees, tasks, and criteria were to be connected in the design such that they could be placed in a common frame of reference (Linacre & Wright, 2002). To generate a connective data matrix, each rater had to provide ratings for the same set of performances, in addition to his or her normal workload. The additional performances, representing the range of TDN levels, had been pre-selected from a larger set of examinee performances in a previous trialling of the respective writing task.

Data analysis

The rating data were analyzed by means of the computer program FACETS (Version 3.66; Linacre, 2010). The program used the ratings that raters awarded to examinees to estimate individual examinee proficiencies, rater severities, task and criterion difficulties, respectively, and scale category difficulties. FACETS calibrated the examinees, raters, tasks, criteria, and the rating scale onto the same equal-interval scale (i.e., the logit scale), creating a single frame of reference for interpreting the results of the analysis (for an introductory overview of MFRM, see Eckes, 2009a, in press).

Variable map

Figure 2 displays the variable map representing the calibrations of the examinees, raters, tasks, criteria, and the four-point TDN rating scale as raters used it to rate examinee speaking performance.

The variability across raters in their level of severity was substantial. Thus, the rater severity measures showed a 2.70-logit spread, which was more than a fifth (21.7%) of the logit spread observed for examinee proficiency measures (12.46 logits). In other words, differences in rater severity were far from being negligible. This was consistently revealed by rater 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 the present group of raters there were about 19 statistically distinct strata of severity (to illustrate, when raters exercised a similar level of severity, an index value close to 1 would be expected), and (c) the rater separation reliability was close to unity, attesting to a very high amount of rater disagreement.

Figure 2. Variable map from the FACETS analysis of TestDaF speaking performance data. Each star in the second column represents 17 examinees, and a dot represents fewer than 17 examinees. The horizontal dashed lines in the last two columns indicate the category threshold measures for the four-category TDN scale (for tasks at TDN 5; i.e., Task 4 and Task 6) and for the three-category TDN scale

(for tasks at TDN 4; i.e., Task 2 and Task 5), respectively; the thresholds for the two-category scale coincide with the difficulty measures of tasks at TDN 3 (i.e., Task 2 and Task 7).

Logit	Examinee	Rater	Task	Criterion	TDN Scales	
	High	***Severe***	***Difficult***	***Hard***	**(TDN 5)**	**(TDN 4)**
7	.					
	.					
	.					
6	.					
	.					
	.					
5	.					
	*.					
	**.					
4	**.					
	****.					
	****.					
3	***.					
	*****.					
	****.				**-----**	
2	*******.					
	*********					
	**********	**	**6**			**-----**
1	*********.	*	**4**		**4**	
	*******.	*****		**standpoint**		
	*****.	******	**3**	**argument. correctness**		
0	***.	**********		**content completeness**	**-----**	**3**
	***.	*****	**5**	**adequacy**		
	**.	*****		**comprehens. vocabulary**		
-1	**.	*	**2**		**3**	
	*.	**				**-----**
	*.		**7**			
-2	*.					
	.				**-----**	
	.					
-3	.					
	.					
	.					
-4	.					
	.					
	.					
-5	.					
	.					
-6	.					
	Low	***Lenient***	***Easy***	***Easy***	**(below 3)**	**(below 3)**

Compensating for rater severity differences

Expressing the rater severity differences in the metric of the TDN scale showed that the most severe rater provided ratings that were, on average, 0.95 raw-score points lower than those provided by the most lenient rater. That is, the severity difference between these two raters was almost one TDN level. Obviously, then, rater severity differences in the order revealed here can have important consequences for examinees.

A case in point is Examinee 561 (see Table 5). Based on the MFRM model, this examinee had an estimated proficiency of 1.99 logits (*SE* = 0.27); the observed average was 3.27. Rounding the observed average to the next TDN level, the final level awarded would have been TDN 3. By contrast, expressing the examinee's proficiency measure in terms of the TDN metric yielded an average of 3.64. This so-called *fair average* is higher than the observed average because it resulted from compensating for the severity of Rater 15 (severity measure = 0.81 logits, *SE* = 0.04), who had happened to rate this examinee's speaking performance. Again rounding to the next TDN level, the examinee would have been awarded the final level TDN 4, making him or her eligible for university admission. Thus, in the case of Examinee 561, applying the many-facet Rasch analysis would have led to an *upward adjustment* of this examinee's result on the speaking test. The same kind of adjustment would have occurred with Examinee 1170, though one TDN level up the scale.

Conversely, Examinee 335 would have received a *downward adjustment* based on the fair average (2.39; below TDN 3), as opposed to the observed average (2.71; TDN 3). In fact, this examinee's performance had been rated by the most lenient rater in the group (Rater 27, severity = –1.30 logits, *SE* = 0.05). Hence, as compared to the other raters in the group, Rater 27 overestimated the proficiency of Examinee 335. This overestimation was corrected by the proficiency measure (or fair average) provided by the MFRM analysis. Table 5 also shows two examinees (i.e., 515, 1631) whose TDN assignments remained unaffected by the score adjustment.

Table 5 - *Examinee Measurement Results (Illustrative Examples)*

Examinee	Proficiency Measures	*SE*	*N* Ratings	Observed Average	Fair Average
515	3.92	0.56	16	4.75	4.84
1170	3.70	0.44	16	4.44	4.81
561	1.99	0.27	48	3.27	3.64
335	-1.87	0.25	48	2.71	2.39
1631	-1.89	0.44	48	2.12	2.38

Note. Proficiency measures are shown in units of the logit scale. *SE* = Standard error. Fair Averages present examinee proficiency measures in units of the TDN scale (with scores from "2" for the lowest category to "5" for the highest category).

Further findings from the MFRM analysis

The MFRM approach makes available a wide variety of analytic procedures and statistical indicators that help to evaluate the quality of speaking performance ratings in almost any desired detail. Probing into the degree of rater variability and compensating for differences in rater severity illustrate some of the practical benefits that may accrue from using MFRM models – clearly important benefits from the point of view of examinees and other stakeholders. Of course, there is much more to be learned from a MFRM analysis of performance ratings. Due to space restrictions, I only briefly outline some further results of the MFRM analysis relevant for an evaluation of the TestDaF speaking test.

Important information on the overall functioning of the speaking performance assessment is provided by the examinee separation reliability statistic. This statistic indicates how well one can differentiate among the examinees in terms of their levels of proficiency. Usually, performance assessment aims to differentiate among examinees in terms of their proficiency as well as possible. Hence, high examinee separation reliability is the desired goal. For the present data, the MFRM analysis showed that this goal had been achieved (reliability = .96). A formally related, practically useful statistic is the examinee separation, or number of examinee strata, index. This index gives the number of measurably different levels of examinee proficiency. In our sample of examinees, the separation index was 6.83, which suggested that among the 1,771 examinees included in the analysis, there were almost seven statistically distinct classes of examinee proficiency. Note that this finding nicely related to the four-category TestDaF scale. That is, the measurement system worked to produce at least as much reliably different levels of examinee proficiency as the TestDaF speaking section was supposed to differentiate.

For each element of each facet, a MFRM analysis provides *fit indices* showing the degree to which observed ratings match the expected ratings that are generated by the model. Regarding the rater facet, fit indices provide estimates of the consistency with which each individual rater made use of the scale categories across examinees, tasks, and criteria. In the present analysis, rater fit indices showed that the vast majority of raters provided highly consistent ratings. Raters exhibiting a tendency toward inconsistency can be subjected to specific rater training in order to reduce this kind of variability (see, e.g., Elder, Knoch, Barkhuizen & von Randow, 2005; Weigle, 1998; Wigglesworth, 1993).

Another point of interest concerns the relative difficulty of the individual speaking tasks. Considering the design of the TestDaF speaking test, Tasks 2 and 7 are supposed to aim at proficiency level TDN 3, Tasks 3 and 5 at TDN 4, and Tasks 4 and 6 at TDN 5. The task measurement results indicated that these three groups of tasks were nicely ordered from less difficult to highly difficult. At the same time, however, difficulty measures *within* two of these groups (i.e., TDN 3 and TDN 4 tasks, respectively) differed significantly by about half a logít (see also Figure 2). Ideally, there should be no significant

difference between tasks aiming at the same TDN. Hence, a finding such as this one needs to be addressed in discussions with item writers in order to come closer to the intended difficulty distribution.

The measurement results for the criterion facet showed that *standpoint* was the most difficult criterion; that is, examinees were less likely to receive a high rating on this criterion than on any of the other criteria; at the opposite end of the difficulty scale were *comprehensibility* and *vocabulary*, which proved to be the easiest ones (see also Figure 2). The criterion difficulty measures showed a 1.23-logit spread, which was quite in line with what could be expected in the present assessment context. Importantly, fit indices for each of the eight criteria stayed well within even narrow quality control limits, attesting to psychometric unidimensionality of this set of criteria (Henning, 1992; McNamara, 1996). That is, all criteria seemed to relate to the same latent dimension, as assumed by the Rasch model used here.

Since the input data to the MFRM analysis were ratings provided on an ordinal scale, the question arises as to how well the categories on the TDN scale, that is, the scores awarded to examinees, are separated from one another. The analysis typically provides a number of useful indices for studying the functioning of rating scales. For example, for each rating scale category, the average of the examinee proficiency measures that went into the calculation of the category calibration measure should advance monotonically with categories. When this pattern is borne out in the data, the results suggest that examinees with higher ratings are indeed exhibiting “more” of the variable that is being measured than examinees with lower ratings. In the present analysis, the findings strongly confirmed that the TDN rating scale categories were properly ordered and working as intended.

More Validity Evidence

As shown above, the overall functioning of the present speaking performance assessment as assessed by means of the examinee separation reliability was highly satisfactory. Computing this statistic across all TestDaF speaking tests administered so far revealed that in not a single case the examinee separation reliability fell below .94. Corroborating this finding, the number of strata index computed on the same data basis ranged from 5.5 to 6.5. Hence, each and every live exam reliably differentiated at least five classes of examinees in terms of their level of speaking proficiency – a clear indication scoring validity (Weir, 2005).

Additional evidence of validity was provided by a benchmarking study (Kecker & Eckes, in press) that followed the empirical approach as suggested by the CEFR manual (Council of Europe, 2003). In this study, experts rated each of nine spoken production samples taken from a live TestDaF exam on a nine-category CEFR scale covering the levels A1 to C2 (including plus levels). Ratings were analyzed

using the FACETS program. The results confirmed that the pre-assigned TDN levels and the intended CEFR levels were in close agreement, except for two samples that appeared to be placed too low on the CEFR scale.

In a related validation study, Kecker and Eckes (in press) examined the extent to which TDN levels that examinees achieved in a live TestDaF speaking test corresponded to the same examinees' CEFR levels awarded to them by their teachers using the global CEFR scale. Cross-tabulation of teacher-assigned CEFR levels and TDN levels revealed that in 72.4% of the cases the levels were as predicted; that is, almost three out of four examinees received TDN levels that were in line with the expected B2–C1 range along the CEFR scale.

Summary and Conclusion

Since its inception in 2001, the TestDaF adaptation of the SOPI to the German academic context has been successful in terms of acceptance by stakeholders, ease and efficiency of administration, and scoring validity. Over the years, tape-mediated test delivery has been increasingly replaced by computer-assisted test administration. This technological advance has contributed to the spreading of the TestDaF in the world, and it has also contributed (in combination with the top-down rating procedure) to the improvement of scoring efficiency.

Concomitant analysis and evaluation of the ratings of examinee speaking performance has shown that, overall, within-rater consistency is sufficiently high. Hence, ongoing rater training and rater monitoring activities have worked out in this respect. However, as evidenced on every occasion by many-facet Rasch analysis of the rating data, between-rater differences in severity have remained at a level that is much too high to be ignored. In order to compensate for severity differences, final TDN levels are assigned to examinees based on the computation of fair averages.

Currently, detailed qualitative and quantitative feedback from examinees, language teachers, and exam officers has stimulated thinking about further improvements that may be achieved in the future. Some issues that figure prominently in this process concern the following questions: (a) Are the fixed planning and response times adequately designed? (b) Are the required speech acts sufficiently distinct and relevant to contemporary academic context? (c) Are the scoring criteria and performance-level descriptors clearly defined and well-differentiated from one another? (d) Are the less-challenging tasks that aim at level TDN 3 really located at the intended CEFR level (i.e., B2)?

Recently, the Center for Applied Linguistics has developed an approach to assessing speaking proficiency that is more flexible than the SOPI. This approach utilizes the Computerized Oral Proficiency Instrument (COPI; Malabonga, Kenyon & Carpenter, 2005; see also Kenyon & Malone, this volume). The COPI allows examinee control over several aspects of test administration, including

control of the time they take to prepare for and respond to a COPI task. Therefore, at least the first question mentioned above may be readily addressed by following the lines laid out by the COPI.

Acknowledgements

I would like to thank my colleagues at the TestDaF Institute for many stimulating discussions on issues concerning the evaluation of the TestDaF speaking test. Special thanks go to Gabriele Kecker for helpful comments on an earlier version of this paper.

References

Attali, Y. (2007). *Construct validity of* e-rater *in scoring TOEFL essays* (Research Report, RR-07-21). Princeton, NJ: Educational Testing Service.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*, 453–476.

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

Bernstein, J., & Cheng, J. (2007). Logic and validation of fully automatic spoken English test. In M. Holland & F. P. Fisher (Eds.), *The path of speech technologies in computer assisted language learning: From research toward practice* (pp. 174–194). Florence, KY: Routledge.

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt, Germany: Lang.

Chapelle, C. A., & Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*. Advance online publication. doi: 10.1177/0265532210364405

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

Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEF)*. Manual (preliminary pilot version). Strasbourg: Language Policy Division.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*, 197–221.

Eckes, T. (2008a). Assuring the quality of TestDaF examinations: A psychometric modeling approach. In L. Taylor & C. J. Weir (Eds.), *Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference May 2005* (pp. 157–178). Cambridge, UK: Cambridge University Press.

Eckes, T. (2008b). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155–185.

Eckes, T. (2009a). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the*

manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H). Strasbourg, France: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/manuel1_EN.asp?#P19_2121

Eckes, T. (2009b). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt/Main, Germany: Lang.

Eckes, T. (in press). *Many-facet Rasch measurement: An introduction*. Frankfurt/Main, Germany: Lang.

Eckes, T., Ellis, M., Kalnberzina, V., Pižorn, K., Springer, C., Szollás, K., & Tsagari, C. (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.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, *2*, 175–196.

Elliot, S. (2003). IntelliMetric™: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, NJ: Erlbaum.

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, UK: Cambridge University Press.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*, 1–11.

Hoyt, W. T., & Kerns, M.-D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*, 403–424.

Kecker, G., & Eckes, T. (in press). Putting the Manual to the test: The TestDaF–CEFR linking project. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft Manual.* Cambridge, UK: Cambridge University Press.

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.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26–43.

Kuo, J., & Jiang, X. (1997). Assessing the assessments: The OPI and the SOPI. *Foreign Language*

Annals, 30, 503–512.

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

Linacre, J. M. (2010). *Facets Rasch model computer program* [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. (2005). *Assessing second language writing: The rater's perspective.* Frankfurt, Germany: Lang.

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

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Malabonga, V., Kenyon, D. M., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing,* 22, 59–92.

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

O'Loughlin, K. J. (2001). *The equivalence of direct and semi-direct speaking tests*. Cambridge, UK: Cambridge University Press.

O'Sullivan, B. (2008). *Modelling performance in tests of spoken language.* Frankfurt, Germany: Lang.

O'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS Writing Module. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 446–478). Cambridge, UK: Cambridge University Press.

Pearson. (2009). *PTE Academic automated scoring.* Retrieved from http://www.pearsonpte.com/SiteCollectionDocuments/AutomatedScoringUK.pdf

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6,* 113–125.

Shermis, M. D., & Burstein, J. C. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11,* 99–123.

Stansfield, C. W., & Kenyon, D. M. (1988). *Development of the Portuguese speaking test*. Washington, DC: Center for Applied Linguistics.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20,* 347–364.

TestDaF Institute. (2010). *Jahresbericht 2008/09* [Annual report 2008/09]. Hagen, Germany: Author.

Wang, J., & Brown, M. S. (2007). Automated essay scoring versus human scoring: A comparative study. *Journal of Technology, Learning, and Assessment, 6,* 4–28.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263–287.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*. Advance online publication. doi: 10.1177/0265532210364406

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

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

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

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater*^{SM} *v1.0* (Research Report, RR-08-62). Princeton, NJ: Educational Testing Service.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English**.** *Speech Communication*, *51*, 883–895.

The author:
Thomas Eckes
TestDaF Institute
Feithstraße 188
D - 58084 Hagen
E-mail: thomas.eckes@testdaf.de

Thomas Eckes has been at the TestDaF Institute, Hagen, Germany, since 2001. He is currently Deputy Director and Head of the Language Testing Methodology, Research, and Validation unit. He has extensive teaching experience in educational measurement and statistics, as well as in cognitive and social psychology. He has published numerous articles in edited volumes and peer-review journals, including *Language Testing*, *Language Assessment Quarterly*, *Diagnostica*, *Journal of Classification*, *Multivariate Behavioral Research*, and *Journal of Personality and Social Psychology*. Recently, he has contributed a chapter on many-facet Rasch measurement to the CEFR Manual Reference Supplement, available online at www.coe.int/lang. His research interests include: rater cognition and rater behaviour; rater effects; polytomous IRT models; construct validation of C-tests; standard setting; computerized item-banking; Internet-delivered testing.