Thomas Eckes, Hagen

Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen

1. Einleitung

Die Beurteilung sprachlicher Leistungen anhand mehrstufiger Skalen mit geordneten Antwortkategorien, wie etwa anhand der bekannten sechsstufigen Notenskala oder verwandter Ratingskalen, ist gängige Praxis von Sprachprüfungen in zahlreichen schulischen und außerschulischen Kontexten. Anders als bei objektiven Tests, bei denen die Fähigkeit der Testpersonen über die Anzahl richtig gelöster Aufgaben gemessen wird, erfolgt bei Leistungsbeurteilungen die Feststellung der Fähigkeit durch *Beurteiler* (oder *Prüfer*, *Korrektoren*), die die beobachteten Leistungen auf einer vorgegebenen Skala einstufen.[1]

Beurteilungsverfahren zur Abschätzung der Sprachfähigkeit oder zur Messung des Sprachstandes im Rahmen von standardisierten Tests haben in den letzten zehn bis fünfzehn Jahren zunehmend an Bedeutung gewonnen (Khattri/Sweet, 1996). Beispielhaft für diese Entwicklung sind die auf breiter Basis durchgeführten nationalen und internationalen Schulleistungsvergleiche. So enthält z.B. der Lesekompetenztest, der in der PISA 2000-Studie (Baumert et al., 2001) eingesetzt wurde, knapp zur Hälfte Aufgaben mit frei zu formulierenden Antworten. Während bei den Mehrfachwahlaufgaben dieses Tests eine Auswertung nach „richtig“ oder „falsch“ erfolgte, hatten bei den offenen Aufgaben Beurteiler eine Bewertung der Leistungen auf der Grundlage eines umfangreichen Katalogs präzise beschriebener Kriterien vorzunehmen (Artelt/Stanat/Schneider/Schiefele, 2001).

Ein anderes prominentes Beispiel für den systematischen Einsatz von Beurteilern sind weltweit durchgeführte Fremdsprachentests wie das „International English Language Testing System“ (IELTS; siehe unter www.ielts.org) oder der „Test Deutsch als Fremdsprache“ (TestDaF; siehe unter www.testdaf.de). Im Falle des TestDaF bewerten Beurteiler die sprachlichen Leistungen in den beiden Subtests Schriftlicher Ausdruck und Mündlicher Ausdruck nach einem detailliert ausgearbeiteten Kriterienkatalog. Im Schriftlichen Ausdruck handelt

[1] Aus Gründen der sprachlichen Vereinfachung werden in dieser Arbeit Ausdrücke wie „Beurteiler“, „Prüfungsteilnehmer“, „Proband“ usw. im generischen Sinne verwendet.

es sich um die Bearbeitung einer einzigen Aufgabe, die nach definierten Einzelkriterien bzw. Deskriptoren schrittweise bewertet wird. Im Mündlichen Ausdruck sind entsprechend die Leistungen bei (maximal) neun Aufgaben zu bewerten.

Ihrer großen Beliebtheit und weiten Verbreitung zum Trotz sind Beurteilungs- oder Ratingverfahren zur Leistungsmessung mit einer Reihe von *Urteilsfehlern* behaftet (vgl. z.B. Bortz/Döring, 2002; Guilford, 1954; Saal/Downey/Lahey, 1980). Zwar sind schon lange vielfältige Formen von Fehlereinflüssen bekannt, aber diese erfahren nur allzu selten eine angemessene Behandlung. Ein zentrales Problem der Leistungsmessung unter Verwendung von Ratingskalen stellt die in vielen Fällen unzureichende Übereinstimmung zwischen den Beurteilern dar. Der vorliegende Beitrag verfolgt das Ziel, *Beurteilungen sprachlicher Leistungen* auf den Prüfstand zu stellen. Am Beispiel der TestDaF-Teilprüfung Schriftlicher Ausdruck soll die grundsätzliche Problematik besprochen, aber auch ein möglicher Weg zur Problemlösung aufgezeigt werden.

Dazu wird wie folgt vorgegangen. Zunächst wird das Problem mangelnder Beurteilerübereinstimmung an einem konkreten Beispiel aus dem schulischen Kontext illustriert. Es wird argumentiert, dass es sich um ein generelles Problem der Leistungsbeurteilung handelt. Als eine wesentliche Ursache geringer Übereinstimmung wird die von Beurteiler zu Beurteiler unterschiedlich ausgeprägte Tendenz zur Strenge bzw. Milde identifiziert. In einem weiteren Abschnitt werden Grundzüge eines testtheoretischen Modells vorgestellt, das es erlaubt, Unterschiede in der Beurteilerstrenge zu messen und bei der Zuweisung von TestDaF-Niveaustufen (TDN-Stufen) zu kontrollieren. Dabei kommt auch die Konsistenz von Leistungsbeurteilungen zur Sprache. Folgerungen aus dem vorgeschlagenen Modell münden in die Entwicklung eines Korrekturverfahrens, dessen besondere Vorzüge an einer Reihe von Beispielen erläutert werden. Abschließend werden Perspektiven einer wissenschaftlich fundierten Leistungsbeurteilung, die eine möglichst objektive und zugleich faire Fähigkeitsmessung zum Ziel hat, diskutiert.

2. Das Problem geringer Beurteilerübereinstimmung

2.1. Benotung von Aufsätzen in der Grundschule

Vor knapp 40 Jahren veröffentlichte der Unterrichtsforscher Rudolf Weiss Ergebnisse einer Untersuchung zur Aufsatzbeurteilung (Weiss, 1965). Sein Untersuchungsansatz war denkbar einfach: Er legte ein und dieselben Aufsätze verschiedenen Lehrern zur Benotung vor. Das Resultat war eindeutig: Die vergebenen Noten streuten enorm stark, in einigen Fällen schöpfte die Streu-

ung der Noten fast die ganze Notenskala aus. Seit Weiss' Aufsehen erregenden Befunden wurde viel über Möglichkeiten einer Verbesserung der offenkundig ungenauen Aufsatzbeurteilung diskutiert, insbesondere wurden Beurteilungskriterien vorgeschlagen, deren Beachtung den Grad der Genauigkeit steigern sollte. Zwar konnten diese Bemühungen einige durchaus beachtliche Erfolge verzeichnen (Lehmann, 1988, 1990), doch blieb insgesamt betrachtet der Gewinn an Urteilsgenauigkeit hinter den Erwartungen zurück. Zudem hatte die Diskussion um eine Objektivierung der Aufsatzbeurteilung stark akademische Züge, mit der Folge, dass empirisch aufwändig geprüfte Kriterien von Lehrern, insbesondere von Grundschullehrern, in der täglichen Beurteilungspraxis kaum oder gar keine Beachtung fanden (vgl. Ingenkamp, 1989, 1995).

Würde sich nach all den Jahren etwas an der Ungenauigkeit von Aufsatzbeurteilungen geändert haben? Zur Beantwortung dieser Frage führten Birkel/Birkel (2002) eine Replikation der Weiss-Studie durch, konsequenterweise ohne Vorgabe von Kriterien. Vier Aufsätze unterschiedlicher Qualität, geschrieben von Schülerinnen und Schülern einer vierten Grundschulklasse, wurden Grundschullehrern zur Benotung vorgelegt. Die Ergebnisse waren wieder eindeutig: Es resultierte eine Verteilung der Noten bei ein und demselben Aufsatz, die fast über die ganze Notenskala reichte. Zur Illustration ist die Notenverteilung für den dritten Aufsatz in Abbildung 1 wiedergegeben.

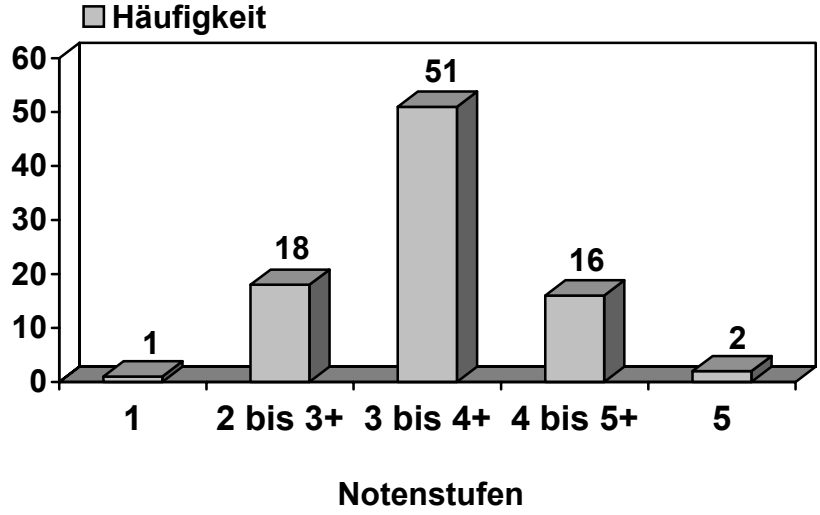

Abbildung 1
Benotung ein und desselben Aufsatzes durch 88 Grundschullehrer (nach Birkel/Birkel, 2002: 222)

Die Autoren kamen zu dem Schluss, dass von Übereinstimmung der Lehrerurteile keine Rede sein könne, und dass die von Weiss (1965) aufgewiesenen Mängel nach wie vor wirksam seien (Birkel/Birkel, 2002: 222).

2.2. Das generelle Problem der Urteilsfehler

Wenn dem so ist, stellt sich die Frage, ob dies nur ein spezielles Problem bei der Aufsatzbeurteilung, vielleicht sogar ein noch spezielleres Problem bei der Aufsatzbeurteilung durch Grundschullehrer ist, oder ob es sich um ein grundsätzliches, auch andere Bereiche betreffendes Problem handelt.

Im Hinblick auf diese Frage sehr aufschlussreich sind Ergebnisse einer *Metaanalyse* von Hoyt/Kerns (1999).[2] Die Autoren betrachteten nicht weniger als 79 unabhängige Einzelstudien, die in ihrer Gesamtheit eine große Bandbreite von beurteilten Merkmalen abdeckten. Neben Leistungsbeurteilungen im engeren Sinne (z.B. Aufsatzbewertungen, direkte Leistungsbeobachtungen, Lehrevaluationen) wurden auch Beurteilungen des sozialen Engagements, der emotionalen Reife und der Qualität dyadischer Interaktionen berücksichtigt. Es zeigte sich, dass im Durchschnitt 37% der Urteilsvarianz auf Urteilsfehler zurückgingen. Waren die Merkmale der direkten Beobachtung nicht zugänglich (z.B. Persönlichkeitseigenschaften oder Fähigkeitsdimensionen, im Unterschied etwa zu Häufigkeitszählungen direkt beobachtbarer Verhaltensweisen), dann belief sich der Anteil der Urteilsfehler an der gesamten Varianz in den Ratingdaten gar auf 49%. Daraus kann der Schluss gezogen werden, dass in den für die Beurteilung sprachlicher Leistungen relevanten Merkmalen etwa die Hälfte der Variabilität in den Urteilen auf Urteilsfehler zurückgeht. Ganz offensichtlich ist das Problem unzureichender Übereinstimmung zwischen Beurteilern nicht nur als generell, sondern auch als substanziell einzustufen.

Doch was genau sind Urteilsfehler? Nach Hoyt (2000) ist unter einem Urteilsfehler („rater bias“) allgemein eine mangelnde Übereinstimmung zwischen Beurteilern zu verstehen, die (a) auf unterschiedliche Interpretationen der Ratingskala oder (b) auf beurteilerspezifische Wahrnehmungen der Beurteilungsobjekte zurückgeht. Beurteilungen liefern damit nicht nur Informationen über das zu messende Konstrukt, sondern auch Informationen über Merkmale der Beurteiler selber. Anders ausgedrückt, Urteilsfehler erhöhen die *konstruktirrelevante* Varianz in den Ratingdaten und mindern so ihre Validität (Kane, 2001; Messick, 1989).

---

[2] Eine Metaanalyse ist eine Art „quantitative Literaturübersicht“, d.h. es werden die Ergebnisse empirischer Einzeluntersuchungen zu einem bestimmten Thema mittels mathematisch-statistischer Methoden quantitativ zusammengefasst (vgl. z.B. Bortz/Döring, 2002, Kap. 9.4).

2.3. Beurteilerübereinstimmung in der TestDaF-Teilprüfung Schriftlicher Ausdruck

Wie weiter oben schon erwähnt, kann die Einführung präzise formulierter Kriterien einen nicht unwesentlichen Beitrag zur Verminderung der Fehlervarianz leisten. Der in der Metaanalyse von Hoyt/Kerns (1999) nachgewiesene hohe durchschnittliche Fehleranteil könnte so gesehen auch daher rühren, dass wenigstens in einem Teil der betrachteten Studien auf eine systematische Verwendung von Beurteilungskriterien verzichtet wurde. Würde ein detailliert ausgearbeiteter Kriterienkatalog nebst intensiven Schulungen von Beurteilern im konsistenten Gebrauch der Kriterien bei der Beurteilung von schriftlichen Sprachleistungen zu einer zufriedenstellenden Urteilsgenauigkeit verhelfen? Diese Frage soll im Folgenden anhand der TestDaF-Teilprüfung Schriftlicher Ausdruck beantwortet werden. Dazu wird auf Leistungsdaten aus einer TestDaF-Prüfung, die im Oktober 2001 weltweit durchgeführt worden war (TestDaF-Prüfung T002), zurückgegriffen.

Ziel des TestDaF-Subtests Schriftlicher Ausdruck ist es festzustellen, inwieweit die Prüfungsteilnehmer in der Lage sind, einen zusammenhängenden und gegliederten Text zu einem hochschulbezogenen Thema zu schreiben. Insbesondere wird die Fähigkeit geprüft, präzise und strukturierte Beschreibungen zu geben und Argumentationen zu entwickeln. Als Vorgabe dienen statistische Diagramme bzw. Tabellen oder schematische Darstellungen von Abläufen sowie Thesen, Fragen oder Zitate. Beispielsweise ist ein Balkendiagramm zur Frage von Studiengebühren in verschiedenen europäischen Ländern vorgegeben. Die Inhalte dieses Diagramms sind zu beschreiben und es ist zu zwei konträren Meinungen über die Einführung von Studiengebühren in Deutschland Stellung zu nehmen.[3] Ungefähr 20 Minuten sind für die Beschreibung und weitere 40 Minuten für die Argumentation aufzuwenden.

Die Beurteilung erfolgt auf der Grundlage klar definierter Kriterien. Die drei übergeordneten Kriterien sind: (a) „Gesamteindruck“ (globale Bewertung des Textes in seiner Wirkung auf den Rezipienten), (b) „Behandlung der Aufgabe“ (Bewertung der Ausführlichkeit und Komplexität der Aufgabenbearbeitung) und (c) „sprachliche Realisierung“ (Bewertung der verwendeten sprachlichen Mittel, vor allem hinsichtlich Breite, Korrektheit und Angemessenheit).

Zu jedem Kriterium existieren drei Unterkriterien; diese lauten z.B. für den Gesamteindruck „Lesefluss“, „Gedankengang“ und „Textstruktur“. Die Un-

[3] Eine komplette Aufgabe dieses Typs findet sich im Modellsatz 01. Dieser kann auf der Internetseite des TestDaF-Instituts (www.testdaf.de/html/modellsatz/sa) eingesehen werden.

terkriterien sind auf jeder TDN-Stufe mit sog. Deskriptoren verbunden, die jeweils Teilaspekte der schriftlichen Ausdrucksfähigkeit erfassen. Beispiele für Deskriptoren auf der Stufe TDN 5 sind: „Der Text liest sich durchgängig flüssig“ (Gesamteindruck), „Die Informationen der Grafik(en) werden zusammengefasst; sie werden klar und folgerichtig dargestellt“ (Behandlung der Aufgabe), „Der Text hat ein breites Spektrum an syntaktischen Strukturen“ (sprachliche Realisierung).

In der TestDaF-Prüfung T002 bewerteten insgesamt 18 Beurteiler die schriftlichen Leistungen von 402 Prüfungsteilnehmern (im Folgenden auch Probanden oder kurz Pbn). Alle Beurteiler verfügten über eine mehrjährige Berufserfahrung im Bereich Deutsch als Fremdsprache. Im Rahmen von universitären oder universitätsnahen Instituten bzw. Organisationen (z.B. Carl-Duisberg-Centren, Studienkollegs) führten sie Sprachkurse hauptsächlich auf Mittel- und Oberstufenniveau durch und waren in der Mehrzahl als Prüfer tätig. In zweitägigen Schulungen wurden die Beurteiler von Mitarbeiterinnen des TestDaF-Instituts auf ihre spezifische Aufgabe im Rahmen des Subtests Schriftlicher Ausdruck gründlich vorbereitet. Die Leistungsbeurteilungen selber folgten einem detailliert ausgearbeiteten Kriterienkatalog. Jede schriftliche Arbeit wurde von zwei unabhängigen Beurteilern auf der TDN-Skala eingestuft. Im Falle einer Nichtübereinstimmung der Gesamtbewertungen wurde eine Drittkorrektur zur endgültigen Festlegung der TDN-Stufe vorgenommen.[4]

In Tabelle 1 finden sich die Ergebnisse zur prozentualen Übereinstimmung zwischen je zwei Beurteilern.[5] Eingang in die Berechnungen fanden die Gesamtbewertungen im Schriftlichen Ausdruck. Als prozentuale Übereinstimmung wurde der Anteil der übereinstimmenden TDN-Stufenzuweisungen an der Gesamtzahl der bewerteten Leistungen ermittelt. Beurteiler mit deutlich weniger als 20 Bewertungen blieben von der Betrachtung ausgeschlossen.

[4] Der TestDaF weist das in jedem der vier Leistungsbereiche (Leseverstehen, Hörverstehen, Schriftlicher Ausdruck, Mündlicher Ausdruck) erzielte Sprachniveau separat anhand der TestDaF-Niveaustufen-Skala (TDN-Skala) aus. Die Charakterisierung der Prüfungsleistungen erfolgt mittels der Stufen „unter TDN 3“, „TDN 3“, „TDN 4“ und „TDN 5“. Dabei entsprechen die TDN-Stufen 3, 4 und 5 den Stufen B2.1 („selbstständige Sprachverwendung“) bis C1.2 („kompetente Sprachverwendung“) des „Gemeinsamen europäischen Referenzrahmens für Sprachen“ (Europarat, 2001, Kap. 3; vgl. auch Quetz, 2003). Unterhalb von TDN 3 erfolgt keine Differenzierung; es wird nur festgestellt, dass das Eingangsniveau von TestDaF noch nicht erreicht ist.

[5] Beurteiler- bzw. Pbn-Kennungen wurden aus Gründen des Datenschutzes verändert.

Tabelle 1.
Prozentuale Übereinstimmungsraten und Interraterreliabilitäten in der TestDaF-Prüfung T002 (Schriftlicher Ausdruck)

Beurteiler	Anzahl beurteilter Personen	Prozentuale Übereinstimmung	Interraterreliabilität
07 – 10	20	70%	**.66**
13 – 16	21	57%	**.65**
12 – 03	20	55%	**.31**
17 – 11	19	53%	**.42**
14 – 08	23	52%	**.50**
08 – 12	24	50%	**.53**
09 – 17	28	46%	**.26**
05 – 18	22	45%	**.50**
02 – 04	27	44%	**.32**
10 – 09	20	40%	**.39**
15 – 07	27	33%	**.16**
13 – 03	21	24%	**.21**
05 – 07	20	20%	**.22**
01 – 14	18	11%	**–.03**

Anmerkung. Leistungsbewertungen erfolgten durch je zwei unabhängige Beurteiler. Die Interraterreliabilität wurde ermittelt nach dem gewichteten Kappa-Koeffizienten (Cohen, 1968). Kappa ist ein zufallskorrigiertes Maß der Beurteilerübereinstimmung für geordnete Kategorien.

Die dritte Spalte gibt die prozentuale Übereinstimmung wieder. Diese reicht von 70% für das Beurteiler-Paar 07–10 bis 11% für das Paar 01–14. Im Durchschnitt belief sich die prozentuale Übereinstimmung auf nur 44%, d.h. in 56% der Fälle musste eine Drittkorrektur über die endgültige TDN-Stufe entscheiden.

Die vierte Spalte enthält die Ergebnisse für die *Interraterreliabilität* als Maß der Genauigkeit der Bewertungen. Bestimmt wurde dieses Maß anhand des (gewichteten) Kappa-Koeffizienten (Cohen, 1968). Das gewichtete Kappa ist ein zufallskorrigierter Koeffizient, der den Grad der Übereinstimmung zwischen zwei Beurteilern bei geordneten Urteilskategorien angibt. Dieser Koeffizient kann Werte zwischen –1 und +1 annehmen. Der maximale Wert wird erreicht, wenn perfekte Übereinstimmung vorliegt (d.h. alle Beurteilten werden in übereinstimmender Weise den TDN-Stufen zugewiesen). Bei Unabhängigkeit der Einstufungen ist Kappa gleich (oder nahe) 0, im Fall systematisch gegensätzlicher Einstufungen wird Kappa negativ. Um von einer „guten" Übereinstimmung sprechen zu können, sollte Kappa mindestens 0.70 betragen (Bortz/Döring, 2002: 277).

Die höchsten, in Tabelle 1 ausgewiesenen Kappa-Werte betragen 0.66 bzw. 0.65; bei den meisten Paaren liegen die zufallskorrigierten Übereinstimmungsmaße unter 0.50. In einem Fall (01–14) ist die Interraterreliabilität sogar schwach negativ. Damit bewegen sich sowohl die Werte für die prozentuale Übereinstimmung als auch die Interraterreliabilitäten in der Mehrzahl der Vergleiche auf einem derart niedrigen Niveau, dass insgesamt nur von einer unbefriedigenden Übereinstimmung gesprochen werden kann.

3. Die Tendenz zur Strenge bzw. Milde

Um den möglichen Ursachen der Nichtübereinstimmung zwischen den Beurteilern nachzugehen, sollen im Folgenden die Bewertungen zweier Beurteiler-Paare im Detail betrachtet werden. Die Bewertungen des Paares 13–16 sind in Tabelle 2 wiedergegeben.

Tabelle 2.
Bewertungen der Beurteiler 13 und 16 in der TestDaF-Prüfung T002 (Schriftlicher Ausdruck)

	Beurteiler 16				
Beurteiler 13	unter TDN 3	TDN 3	TDN 4	TDN 5	Zeilensumme
unter TDN 3	**8**				8
TDN 3	**1**	**1**			2
TDN 4		**5**	**2**	**3**	10
TDN 5				**1**	1
Spaltensumme	9	6	2	4	21

Anmerkung. Die prozentuale Übereinstimmung bei 21 beurteilten Personen beträgt 57% (12 Übereinstimmungen, 9 Nichtübereinstimmungen; Kappa = 0.65).

Mit einer prozentualen Übereinstimmung von 57% und einer Interraterreliabilität von 0.65 lagen für dieses Paar noch relativ gute Werte vor. Die neun Nichtübereinstimmungen betrafen jeweils nur 1 TDN-Stufe. So hatte z.B. Beurteiler 13 fünf Pbn auf TDN 4 eingestuft, aber Beurteiler 16 dieselben Pbn auf TDN 3.

In Tabelle 3 sind wieder die Bewertungen von Beurteiler 13 aufgeführt, dieses Mal aber im Vergleich zu Beurteiler 03.

Tabelle 3.
Bewertungen der Beurteiler 13 und 03 in der TestDaF-Prüfung T002 (Schriftlicher Ausdruck)

	Beurteiler 03				
Beurteiler 13	unter TDN 3	TDN 3	TDN 4	TDN 5	Zeilensumme
unter TDN 3	**1**	**2**	**2**		5
TDN 3			**6**	**4**	10
TDN 4			**2**	**2**	4
TDN 5				**2**	2
Spaltensumme	1	2	10	8	21

Anmerkung. Die prozentuale Übereinstimmung bei 21 beurteilten Personen beträgt 24% (5 Übereinstimmungen, 16 Nichtübereinstimmungen; Kappa = 0.21).

Für das Paar 13–03 ergab sich eine sehr niedrige, aber keineswegs untypische Übereinstimmungsrate von 24% (Kappa = 0.21). Zehn der 16 Nichtübereinstimmungen betrafen 1 TDN-Stufe, die sechs übrigen betrafen 2 TDN-Stufen. So hatte z.B. Beurteiler 13 vier Pbn nach TDN 3 eingestuft, Beurteiler 03 dieselben Pbn aber nach TDN 5. Die Verteilung der nichtübereinstimmenden TDN-Zuweisungen scheint im zweiten Beispiel alles andere als zufällig zu sein. In keinem einzigen Fall hat Beurteiler 03 Pbn niedriger eingestuft als Beurteiler 13. Die Tendenz von 03 geht damit eindeutig in die Richtung einer *systematisch höheren* Einstufung.

Im ersten Beispiel drängt sich die Vermutung auf, dass die beiden Beurteiler ähnlich streng oder ähnlich milde bewerteten; dagegen legt das Bewertungsmuster im zweiten Beispiel nahe, dass Beurteiler 03 deutlich milder war als Beurteiler 13.

Die Strenge (oder Milde) beschreibt allgemein die Tendenz, niedrigere (bzw. höhere) Einstufungen auf der Skala der TestDaF-Niveaustufen vorzunehmen, als es der tatsächlichen Leistung entspricht. Strenge Beurteiler vergeben generell eher niedrigere Bewertungen, d.h., sie tendieren zu einer *Unterschätzung* der Fähigkeit von Pbn entlang des Fähigkeitskontinuums; milde Beurteiler vergeben generell eher höhere Bewertungen, d.h., sie tendieren zu einer *Überschätzung* der Fähigkeit von Pbn entlang des Fähigkeitskontinuums.

Niedrige Übereinstimmungsraten, wie sie im Beispiel verdeutlicht wurden, sind gleich in mehrfacher Hinsicht problematisch. Erstens verweisen sie aus traditioneller testtheoretischer Perspektive auf eine *mangelnde Genauigkeit* der abgegebenen Bewertungen (vgl. z.B. Wirtz/Caspar, 2002). Grundsätzlich wäre aus dieser Sicht zu fordern, dass zwei Beurteiler für dieselbe Leistung, die sie unabhängig voneinander bewerten, ein und dieselbe TDN-Stufe vergeben sollten. Zweitens machen Nichtübereinstimmungen (im Rahmen eines traditionellen Korrekturverfahrens) die Durchführung einer *Drittkorrektur* notwendig, um eine Entscheidung über die zu vergebende TDN-Stufe zu treffen. Dies lässt aber das zugrunde liegende Problem unberührt und ist (insbesondere bei Sprachprüfungen mit vielen Pbn) zeit- und kostenintensiv. Drittens führt geringe Übereinstimmung zwischen Beurteilern in der Regel dazu, dass Anstrengungen unternommen werden, die Übereinstimmungsrate auf ein zufriedenstellendes Niveau anzuheben. Durch (wieder zeit- und kostenintensive) Nachschulungsmaßnahmen soll im traditionellen Ansatz eine möglichst weitgehende *Homogenisierung* der Beurteiler hinsichtlich ihrer Bewertungsstandards erreicht werden.

Ergebnisse der internationalen Sprachtestforschung unterstreichen jedoch, dass Trainings zur Vereinheitlichung von Bewertungsstandards in aller Regel nur wenig Erfolg haben. Beurteiler unterscheiden sich hinsichtlich ihrer Tendenz zur Strenge bzw. Milde stark voneinander, zeigen diese Unterschiede auch über einen Zeitraum von mehreren Jahren und lassen sich selbst durch zeitlich ausgedehnte, intensive Schulungen nur selten zur Anwendung hinreichend ähnlicher Standards bewegen (Eckes, 2004; Engelhard, 2002; McNamara, 1996).

Wie kann dann aber eine Lösung des Problems geringer Übereinstimmungsraten aussehen? Lösungsansätze, die im Rahmen der *Klassischen Testtheorie* (KTT; vgl. z.B. Gulliksen, 1950; Lienert/Raatz, 1998; Wirtz/Caspar, 2002) entwickelt wurden, teilen die Vorstellung von einem *idealen Beurteiler*, der mit seinem Urteil exakt den wahren Leistungs- oder Fähigkeitswert („True Score“) einer Person trifft. Ganz im Sinne einer solchen *True-Score-Konzeption* sehen es Wirtz/Caspar (2002: 15) als einen zentralen Aspekt der Güte von Beurteilungen an, dass die Beurteiler prinzipiell *austauschbar* sind, d.h., die Unterschiede zwischen den Urteilen verschiedener Beurteiler, die dieselbe Person einstufen, sollten vernachlässigbar klein sein. Erst wenn das Kriterium der Austauschbarkeit (hinreichend) erfüllt sei, so Wirtz/Caspar, könne von einer hohen Präzision der Urteile eines gegebenen Beurteilers gesprochen werden. Wie oben argumentiert wurde, ist es jedoch in den meisten Beurteilungskontexten schlicht illusorisch, dieses Kriterium auch nur annähernd zu erreichen.

Eine Lösung ist dagegen möglich, wenn man die Fehleranfälligkeit von Beurteilungen aus der Perspektive der *Item-Response-Theorie* (IRT; vgl. z.B. Embretson/Reise, 2000; Hambleton/Robin/Xing, 2000; Rost, 2004) betrachtet. Mit der *Multifacetten-Rasch-Analyse* („many-facet Rasch measurement“; Linacre, 1989; Linacre/ Wright, 2002) wird im nächsten Abschnitt ein spezielles IRT-Modell in seinen Grundzügen vorgestellt, das die Beurteilerstrenge zu messen und bei der Festlegung der TDN-Stufen zu berücksichtigen erlaubt (vgl. für eine detailliertere Darstellung Eckes, 2004).

4. Grundzüge der Multifacetten-Rasch-Analyse

4.1. Das Multifacetten-Rasch-Modell

Die Konstruktion und Analyse von Tests sowie ihre Verwendung in der diagnostischen Praxis haben sich in den letzten Jahrzehnten tiefgreifend verändert. Maßgeblichen Anteil daran hat die Pionierarbeit des dänischen Mathematikers Georg Rasch (1960/1980). Rasch hat eine Reihe von probabilistischen

Testmodellen entwickelt, die zusammen mit ihren zahlreichen Weiterentwicklungen eine prominente Rolle innerhalb der Klasse von IRT-Modellen spielen.

Item-Response-Modelle sind, kurz gesagt, formale Modelle des Antwortverhaltens, das Personen bei der Bearbeitung einzelner Testaufgaben (Items) bzw. Beurteiler bei der Einstufung sprachlicher Leistungen zeigen. Das Multifacetten-Rasch-Modell („many-facet Rasch measurement"; kurz MFRM- oder auch *Facets*-Modell; Linacre, 1989; Linacre/Wright, 2002) erlaubt es, Unterschiede in der Beurteilerstrenge zu erfassen und die abgegebenen Urteile entsprechend zu korrigieren. Darüber hinaus teilt es mit anderen Rasch-Modellen den Vorteil einer Konstruktion *linearer Maße* der in Frage stehenden latenten Variablen (z.B. Beurteilerstrenge, Personenfähigkeit, Aufgabenschwierigkeit), ohne (wie es für klassische Ansätze typisch ist) die Schätzungen der jeweiligen Parameter miteinander zu konfundieren. Zudem liefert dieses Modell zu jedem individuellen Parameterwert eine Angabe der Messgenauigkeit durch Schätzung des Standardfehlers.

In testtheoretischer Hinsicht ist festzuhalten, dass das MFRM-Modell das Rasch-Modell für dichotome Antwortvariablen (Rasch, 1960/1980; Wright/Stone, 1979) bzw. Rasch-Modelle für polytome Antwortvariablen mit geordneten Kategorien, insbesondere das Ratingskalen-Modell (Andrich, 1978) und das Partial-Credit-Modell (Masters, 1982), erweitert (vgl. für eine kompakte Darstellung dieser Modelle Eckes, in Druck; Wright/Masters, 1982). Eine MFRM-Analyse kann mit dem von Linacre (1999) entwickelten Programm FACETS durchgeführt werden.

In einer Sprachprüfung wie TestDaF bestimmen folgende (systematische) Faktoren die Leistungsbeurteilung: (a) die *Fähigkeit* der Pbn (leistungsstärkere Pbn sollten höhere Einstufungen, leistungsschwächere Pbn niedrigere Einstufungen erhalten), (b) die *Schwierigkeit* der Kriterien (im Schriftlichen Ausdruck; ein „schwieriges" Kriterium ist ein solches, bei dem die Pbn generell eher niedrigere Einstufungen erhalten) bzw. die Schwierigkeit der Aufgaben (im Mündlichen Ausdruck; eine „schwierige" Aufgabe ist entsprechend eine solche, bei der die Pbn generell eher niedrigere Einstufungen erhalten) und (c) die (bereits erläuterte) *Strenge* (oder *Milde*) der Beurteiler. Entsprechend der hier betrachteten Modellanwendung werden im Folgenden die beurteilten Personen, die Beurteiler und die Beurteilungskriterien als Facetten der Urteilssituation aufgefasst.

Allgemeines Ziel einer Multifacetten-Rasch-Analyse ist es, möglichst objektive und präzise Informationen über die Elemente der betrachteten Facetten zu gewinnen. Sie soll also (im typischen Fall) Aufschluss geben nicht nur über

die Leistungsfähigkeit der beurteilten Personen, sondern gleichzeitig auch über die Strenge der Beurteiler und die Schwierigkeit der Aufgaben bzw. Kriterien. Die quantitativen Aussagen über die Elemente einer bestimmten Facette sollen dabei so weit wie möglich unabhängig von den Aussagen über die Elemente der anderen Facetten sein. Das heißt, die von einer MFRM-Analyse gelieferten Schätzungen der Leistungsfähigkeit der Personen sollen weder von der Verteilung der Strenge der Beurteiler noch von der Verteilung der Schwierigkeit der Kriterien beeinflusst sein. Entsprechendes sollte für die Schätzungen der Beurteilerstrenge und der Kriterienenschwierigkeit gelten.

Das zugrunde gelegte MFRM-Modell hat in logarithmischer Schreibweise die Form (vgl. Linacre, 1989):

$$\ln\left[\frac{p_{vijk}}{p_{vijk-1}}\right] = \theta_v - \beta_i - \alpha_j - \tau_k .$$

Dabei haben die einzelnen Symbole folgende Bedeutung:

p_{vijk} = Wahrscheinlichkeit einer Einstufung von Person v bei Kriterium i durch Beurteiler j in Kategorie k,
p_{vijk-1} = Wahrscheinlichkeit einer Einstufung von Person v bei Kriterium i durch Beurteiler j in Kategorie k – 1,
θ_v = Fähigkeitsparameter von Person v,
β_i = Schwierigkeitsparameter von Kriterium i,
α_j = Strengeparameter von Beurteiler j,
τ_k = Schwierigkeitsparameter von Kategorie k.

Der logarithmierte Quotient links vom Gleichheitszeichen in der Modellgleichung wird *Logit* genannt („ln“ steht für den natürlichen Logarithmus). Danach ist der Logit eine lineare Funktion der Personenfähigkeit θ_v, der Itemschwierigkeit β_i, der Beurteilerstrenge α_j und der Kategorieschwierigkeit τ_k. Die Parameter sind mit Minuszeichen verknüpft, sodass der Beurteilerparameter α_j die Strenge und nicht die Milde des Beurteilers j ausdrückt; entsprechend drückt der Itemparameter β_i die Schwierigkeit und nicht die Leichtigkeit des Items i aus (das Gleiche gilt für den Kategorieparameter τ_k).

Der Schwierigkeitsparameter von Kategorie (bzw. TDN-Stufe) k gibt an, wie wahrscheinlich es ist, eine Einstufung in Kategorie k zu erhalten, relativ zu einer Einstufung in Kategorie k – 1 (d.h., je höher der Parameterwert, desto weniger wahrscheinlich ist eine Einstufung in k). Zugleich definiert dieser

Parameter in der obigen Gleichung, wie die Ratingdaten zu behandeln sind. Im vorliegenden Fall wird das Ratingskalen-Modell von Andrich (1978) spezifiziert. Danach wird bei den Bewertungen auf allen Kriterien die gleiche Struktur der Ratingskala zugrunde gelegt, d.h., die Abstände zwischen den Schwellen[6] werden bei allen Kriterien als gleich groß angenommen.

In der logarithmischen Darstellung des Modells wird deutlich, dass alle Modellparameter auf einer gemeinsamen linearen Skala, d.h. auf der *Logitskala*, kalibriert werden. Dies ist eine für praktische Anwendungen bedeutsame Eigenschaft aller Rasch-Modelle. Die allgemeine Beziehung zwischen der Lösungswahrscheinlichkeit und der Ausprägung der latenten Variablen wird in diesen Modellen durch eine S-förmige *logistische Funktion* beschrieben (vgl. Fischer, 1974; Rost, 2004; Steyer/Eid, 2001).

4.2. Kontrolle der Modellgeltung: Fit-Statistiken

Eine MFRM-Analyse liefert zu jedem Element jeder einzelnen Facette einen Messwert (Logitwert), einen Standardfehler (d.h. Information über die Genauigkeit des Messwerts) und verschiedene Fitwerte (d.h. Information darüber, wie gut die Daten den Erwartungen des Messmodells entsprechen). Das Ausmaß der Abweichungen der beobachteten von den erwarteten Ratings, die als standardisierte Residuen ausgedrückt werden, gibt Hinweise auf Beurteilungen, die Besonderheiten aufweisen und somit genauer zu untersuchen sind. Eine *Residuenanalyse* erlaubt es, die psychometrische Qualität der Ratingdaten im Detail zu überprüfen.

Die standardisierten Residuen lassen sich über verschiedene Facetten und über die verschiedenen Elemente einer gegebenen Facette hinweg zusammenfassen, um die Modellgeltung in allen Teilen des Messmodells zu kontrollieren. In der Regel geschieht dies mittels zweier *Mean-Square-Fehlerstatistiken* (vgl. Wright/Masters, 1982: 99), der Outfit- und der Infit-Statistik (vgl. auch Eckes, 2004).

Die *Outfit-Statistik* („Outfit“ steht für „outlier-sensitive fit“) erfasst primär, inwieweit ein ansonsten konsistent einstufender Beurteiler unerwartete Ratings in den äußeren Skalenbereichen abgibt. Dagegen besitzt die *Infit-Statistik* („Infit“ steht für „inlier-sensitive“ oder auch „information-weighted fit“) eine größere Empfindlichkeit im Falle unerwarteter Ratings, die sich im mittleren Skalenbereich bewegen.

[6] Schwellen lassen sich als diejenigen Punkte auf dem Fähigkeitskontinuum definieren, an denen der Übergang von einer Antwortkategorie der Ratingskala zur nächsten stattfindet. Je höher die Schwelle, desto schwieriger der Übergang.

Infit- und Outfit-Statistik haben einen Erwartungswert von 1; sie können Werte im Bereich zwischen 0 und +∞ annehmen (Linacre, 2003; Wright/Masters, 1982). Fitwerte deutlich größer 1 verweisen darauf, dass die Ratingdaten mehr Variation besitzen, als es den Erwartungen des Modells entspricht. Allgemein liegt bei einem Wert von 1 + x die Variation um 100x% höher als erwartet. Umgekehrt indizieren Fitwerte deutlich kleiner 1, dass die Ratingdaten weniger Variation zeigen als vorhergesagt. Bei einem Wert von 1 – x liegt die Variation um 100x% niedriger als erwartet (vgl. auch Bond/Fox, 2001: 176ff).

Linacre (2002; vgl. auch Wright/Linacre, 1994) hat grobe Richtwerte für die Interpretation von Mean-Square-Statistiken vorgeschlagen. Danach können Infit- bzw. Outfit-Werte im Intervall zwischen 0.5 und 1.5 als messmethodisch akzeptabel gelten.[7]

4.3. Differenziertheit der Leistungsmessung: Separationsstatistiken

Eine Grundvoraussetzung der Nützlichkeit von Leistungsbeurteilungen ist, dass sie hinreichend genau zwischen den beurteilten Personen zu unterscheiden erlauben. FACETS liefert hierzu mit dem Separationsindex, dem Index der Klassenseparation und der Separationsreliabilität drei informative Statistiken. Dabei kommt den letzten beiden besondere praktische Bedeutung zu.

Der *Separationsindex* ist ein Maß für die Streubreite der Leistungsmaße relativ zu ihrer Genauigkeit (Wright/Masters, 1982: 106). Diese Separation wird ausgedrückt als das Verhältnis von „wahrer" Streuung der Leistungsmaße (d.h. der Streuung der Leistungsmaße nach Standardfehlerkorrektur) zum „durchschnittlichen" Standardfehler der Leistungsmaße („Root Mean Square Error"). Im Hinblick auf die Beurteiler gibt der Separationsindex an, wie verlässlich zwischen den Beurteilern anhand ihrer Strengemaße unterschieden werden kann. Analog liefert der Separationsindex im Falle der Kriterien Information über den Grad ihrer Unterscheidbarkeit anhand der Schwierigkeitsmaße.

Auf der Basis des Separationsindex lässt sich für jede Facette ein *Index der Klassenseparation* berechnen. Dieser Index schätzt die Anzahl der aufgrund der Messwerte potenziell unterscheidbaren Klassen von Elementen einer bestimmten Facette („number of strata" bei Wright/Masters, 1982, 2002). So würde z.B. ein Wert der Klassenseparation von 2.1 für die Facette der Beurteiler besagen, dass die Messung es erlaubt, die Gesamtgruppe der Beurteiler in

[7] Je nach Fragestellung oder Verwendungszusammenhang der Untersuchungsergebnisse können die Intervalle auch breiter oder enger definiert werden (vgl. Bond/Fox, 2001: 176ff).

ca. zwei statistisch zuverlässig unterscheidbare Klassen einzuteilen. Ein Index von ungefähr 1.0 würde dagegen bedeuten, dass die Beurteiler vernachlässigbar geringe Unterschiede in ihren Strengemaßen aufweisen und damit als untereinander austauschbar gelten können.

Die wohl informativste Separationsstatistik ist die *Separationsreliabilität*. Diese lässt sich ausdrücken als Verhältnis von messfehlerkorrigierter Varianz zu beobachteter Varianz. Je nach betrachteter Facette erfährt die Separationsreliabilität eine andere inhaltliche Interpretation. Im Falle der beurteilten Personen ist diese Statistik Cronbachs Alpha vergleichbar, d.h., sie gibt den Grad der Genauigkeit an, mit der Unterscheidungen zwischen den Personen hinsichtlich ihrer Fähigkeit getroffen werden können. Richtet sich der Blick auf die Beurteiler, so zeigt die Separationsreliabilität an, wie sehr sich die einzelnen Strengemaße voneinander unterscheiden. Ganz im Gegensatz zur Interraterreliabilität, die (allgemein gesprochen) ein Index dafür ist, wie *ähnlich* sich die Beurteiler in ihrem Urteilsverhalten sind, wird mit der Separationsreliabilität dieser Facette erfasst, wie *unähnlich* sich die Beurteiler sind. Was die Beurteilungskriterien betrifft, gilt analog, dass ein hoher Wert der Separationsreliabilität auf große Unterschiede in ihren Schwierigkeitsmaßen verweist.

Schließlich erlaubt ein approximativer Chi-Quadrat-Test (Hedges/Olkin, 1985: 123) die Prüfung der Hypothese, dass die jeweiligen Parameterschätzungen aus einer Population mit homogenen Parameterwerten stammen (es handelt sich also um eine Homogenitätsstatistik). Allerdings ist bei dieser Art von Signifikanztest eine hohes Maß an Abhängigkeit vom Stichprobenumfang gegeben. Bei großen Stichproben werden schon kleinste Abweichungen von der Homogenitätsannahme als signifikant ausgewiesen.

4.4. Korrektur der Beurteilerstrenge: Fairer Durchschnitt

Für die Elemente jeder einzelnen Facette werden erwartete Beurteilungen ermittelt, die die Variabilität der betreffenden Maße in Rechnung stellen. Da Leistungsbeurteilungen nicht nur so genau, sondern auch so *fair* wie möglich sein sollten, kommt es darauf an, die beobachteten Einzelurteile so zu korrigieren, dass leistungs- bzw. konstruktirrelevante Faktoren keinen substanziellen Einfluss auf die endgültige Einstufung haben.

Die Beurteilerstrenge ist ein solcher einflussreicher Faktor. Um diesen Faktor soweit wie möglich zu kontrollieren und die abschließende Fähigkeitsschätzung nicht von einer mehr oder weniger willkürlichen Zuweisung von Beurteilern abhängig zu machen, ist für jede beurteilte Person dasjenige Rating zu ermitteln, das zustandekäme, wenn die Person von einem Beurteiler mit durchschnittlicher Strenge beurteilt worden wäre. Dieser hypothetische Beur-

teiler wäre also weder milder noch strenger als die übrigen Beurteiler. Unterbliebe eine solche Korrektur, würden *leistungsirrelevante* Aspekte der Testsituation (eben z.B. die Strenge der Beurteiler) die Messung der Sprachfähigkeit beeinflussen. Sind große Unterschiede in der Tendenz zur Strenge bzw. Milde gegeben (wie häufig der Fall), dann hätten jene Pbn einfach Pech, denen zwei strenge Beurteiler zugeteilt wurden; andere Pbn dagegen könnten gleichsam von Glück reden, wenn zwei milde Beurteiler ihre Leistungen bewerteten. Wenn ein Beurteiler zu großer Strenge neigte, der andere aber zu großer Milde, dann wäre das *vorhersagbare* Ergebnis eine ausgeprägte Diskrepanz in den Bewertungen derselben Prüfungsleistungen. Im traditionellen Korrekturverfahren würde diese Nichtübereinstimmung Drittkorrekturen notwendig machen.

FACETS liefert für jedes einzelne Element jeder Facette eine erwartete Einstufung, die auf der Basis der Durchschnitte der jeweils anderen Facetten berechnet wird. Diese erwartete mittlere Einstufung wird auch *fairer Durchschnitt* genannt. Der faire Durchschnitt für eine Person gibt danach ihre um die Strenge bzw. Milde der involvierten Beurteiler wie auch um die Schwierigkeit des jeweiligen Kriteriums bereinigte mittlere Einstufung in der Metrik der Ratingskala an. Auf die Beurteiler übertragen bedeutet diese statistische Methode der Adjustierung bzw. Korrektur, dass ein mittleres Rating unter Berücksichtigung der Leistungsstärke der beurteilten Personen und der Schwierigkeit der Kriterien bestimmt wird. Dies ermöglicht direkte Vergleiche zwischen den Beurteilern, die unabhängig von den Fähigkeiten der jeweils Beurteilten sind.

5. Beurteilerstrenge und Beurteilerkonsistenz: Eine exemplarische Analyse

5.1. Der Facettenraum

Wie schon ausgeführt, werden im Rahmen einer MFRM-Analyse alle Facetten simultan untersucht und auf derselben linearen Skala (Logitskala) kalibriert. Eine grafische Darstellung der Ergebnisse dieser gemeinsamen Kalibrierung erfolgt im sog. *Facettenraum* (Eckes, 2003, 2004). Der Facettenraum erlaubt Vergleiche zwischen den Elementen einer bestimmten Facette wie auch Vergleiche zwischen den verschiedenen Facetten. Damit vermittelt diese Darstellung einen raschen Überblick über die Messergebnisse und bietet einen gemeinsamen Bezugsrahmen für die Ergebnisinterpretation. Abbildung 2 zeigt den Facettenraum für die vorliegende Analyse.[8]

[8] Um den Ursprung der Logitskala festzulegen, wurden (wie in dieser Art von Rasch-Skalierung üblich) die Beurteiler- und Kriterienfacetten zentriert (vgl. Linacre, 2002a: 31).

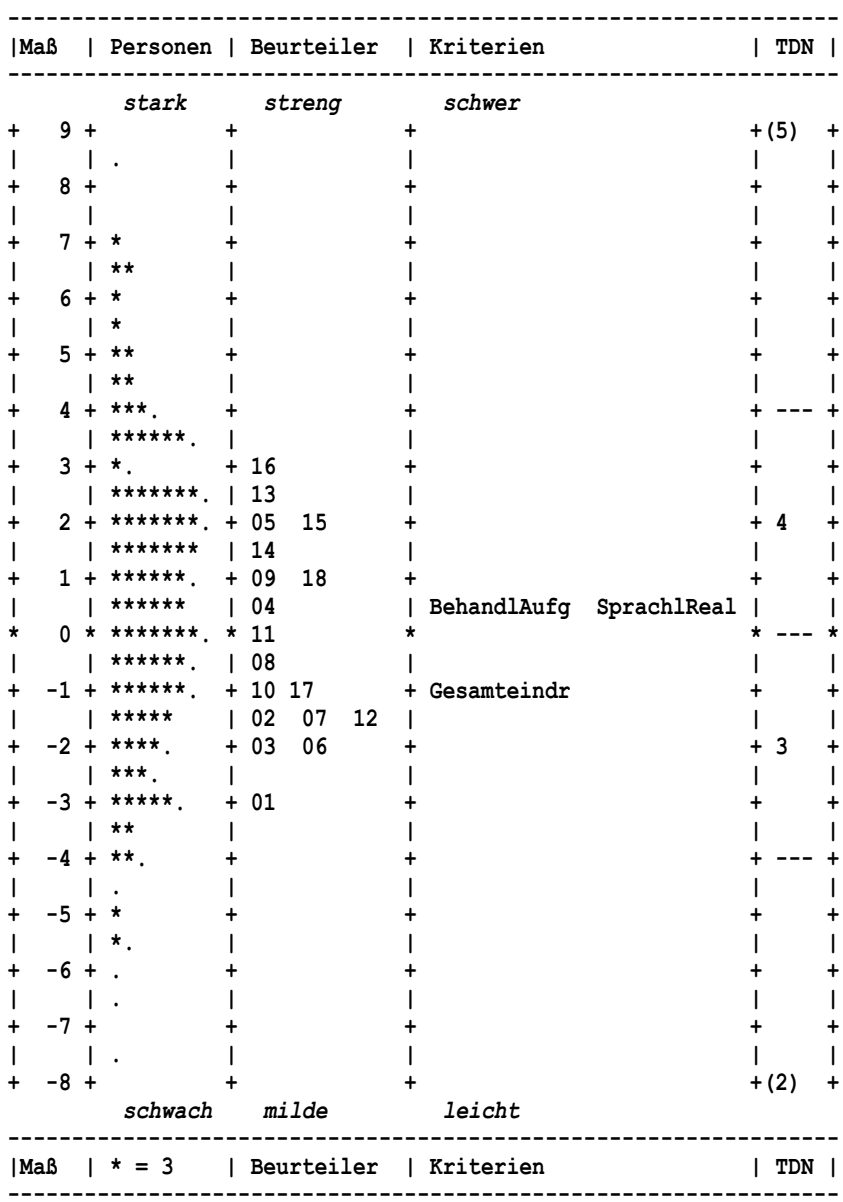

Abbildung 2
Facettenraum zur Beurteilung der Leistungen im schriftlichen Ausdruck

Der Facettenraum ist in fünf Spalten unterteilt. Es seien zunächst die erste und die letzte Spalte betrachtet. Die erste Spalte enthält die Logitskala. Ein Wert auf dieser Skala gibt das Fähigkeitsmaß eines Pb (oder das Strengemaß eines Beurteilers bzw. das Schwierigkeitsmaß eines Kriteriums) wieder. Unterschiede zwischen Pbn, die in Einheiten der Logitskala ausgedrückt werden, haben unabhängig vom Fähigkeitsniveau stets die gleiche Bedeutung. Es handelt sich also um eine lineare Skala.

Die letzte Spalte bildet die TDN-Skala auf die Logitskala ab, d.h., sie gibt die Fähigkeit der Pbn in der Metrik der TDN-Skala wieder. Genauer gesagt, diese Skala repräsentiert diejenigen TDN-Einstufungen, die Pbn mit gegebener Fähigkeit von einem durchschnittlich strengen Beurteiler auf einem durchschnittlich schwierigen Kriterium unter den Modellannahmen erhalten würden.

Die gestrichelten horizontalen Linien in dieser Spalte sind genau an den Schnittpunkten zweier benachbarter Antwortkategorien eingezeichnet. Diese Schnittpunkte (oder Schwellen) lokalisieren diejenigen Punkte auf dem Fähigkeitskontinuum, an denen der Übergang von einer Kategorie zur nächsten stattfindet. Mit anderen Worten, die Wahrscheinlichkeit einer Zuweisung zur nächsthöheren TDN-Stufe wird von diesen Punkten an größer als die Wahrscheinlichkeit einer Zuweisung zur vorherigen TDN-Stufe. Es ist leicht zu erkennen, dass die Schwellen zwischen den Kategorien „unter TDN 3“ und „TDN 3“ sowie zwischen „TDN 3“ und „TDN 4“ etwa äquidistant sind. Dies ist ein Hinweis auf Intervallskalenqualität im mittleren Bereich der TDN-Skala.

Die zweite Spalte zeigt die Verteilung der Parameterschätzungen im Hinblick auf die sprachliche Leistungsfähigkeit der Pbn. Leistungsstärkere Pbn finden sich im Facettenraum weiter oben, leistungsschwächere Pbn weiter unten (in der Abbildung steht jedes Sternchen für 3 Pbn, jeder Punkt für 1 bis 2 Pbn).[9] Die Fähigkeitsmaße reichen von 8.30 Logits am oberen Ende des Leistungsspektrums bis –7.64 Logits am unteren Ende. Damit beträgt ihre Streubreite 15.94 Logits.

In der dritten Spalte ist die Verteilung der Schätzungen für den Strengeparameter der Beurteiler wiedergegeben, wobei strengere Beurteiler weiter oben, mildere Beurteiler weiter unten dargestellt sind. Ganz offenkundig ist die Variabilität innerhalb der Beurteilerfacette substanziell: Die Strengemaße

[9] Pbn mit extremen Einstufungen (d.h. durchgängig „unter TDN 3“ oder durchgängig „TDN 5“) blieben hier wie bei den folgenden Analysen unberücksichtigt, da in solchen Fällen nur approximative Parameterschätzungen möglich sind.

reichen von 3.00 am oberen Ende der Logitskala bis –2.78 am unteren Ende. Mit 5.78 Logits beträgt die Streubreite der Strengemaße 36% der Streubreite der Fähigkeitsmaße. Eine Entscheidung darüber, welche beiden Beurteiler die schriftliche Leistung eines bestimmten Pbn bewerten, muss damit erheblichen Einfluss auf das Bewertungsergebnis nehmen. Aus der Darstellung der Strengemaße im Facettenraum ist auch unmittelbar ersichtlich, dass die in Abschnitt 3 geäußerte Vermutung über die Unterschiede in der Strenge von Beurteiler 13 und 16 einerseits und 13 und 03 andererseits zutrifft: Die ersten beiden sind die strengsten von allen Beurteilern, während Beurteiler 03 einer der mildesten ist. Es ist daher alles andere als eine Überraschung, dass Beurteiler 13 und 16 ähnliche (und zwar ähnlich strenge) Einstufungen abgaben, Beurteiler 13 und 03 dagegen zu weit differierenden Einschätzungen kamen.

In der vierten Spalte sind die drei Kriterien (Gesamteindruck, Behandlung der Aufgabe, sprachliche Realisierung) entsprechend ihrer Schwierigkeitsmaße angeordnet. Offenkundig war es in dieser Prüfung schwieriger, in den (analytischen) Kriterien der Behandlung der Aufgabe bzw. der sprachlichen Realisierung eine gute Beurteilung zu erfahren als im (holistischen) Kriterium des Gesamteindrucks.

5.2. Beurteilerkonsistenz

Eine weitere wichtige Frage, die von einer Multifacetten-Rasch-Analyse beantwortet werden kann, betrifft die Konsistenz innerhalb der Beurteiler. Im vorliegenden Kontext wird von *Konsistenz* dann gesprochen, wenn das Urteilsverhalten eines bestimmten Beurteilers mit den Erwartungen des MFRM-Modells in Einklang steht. Da das Multifacetten-Rasch-Modell die Leistungsbeurteilung als einen stochastischen Prozess konzipiert, wird ein gewisses Maß an zufälliger Schwankung modellimmanent vorausgesetzt. „Perfekte“ Konsistenz im Sinne gleich bleibender Beurteilungen der Leistungen über Pbn und Kriterien hinweg würde danach den Erwartungen des Modells geradezu widersprechen.

Auskunft über den Grad der Konsistenz jedes einzelnen Beurteilers geben die in der vierten und fünften Spalte von Tabelle 4 aufgeführten Werte der Infit- und Outfit-Statistiken. Diese Statistiken fassen das Ausmaß der Schwankungen über alle jeweils beurteilten Pbn und über alle Kriterien zusammen. Sie zeigen an, inwieweit die Bewertungen eines gegebenen Beurteilers mehr oder weniger Variation aufweisen, als vom Modell erwartet wird (auch *interne* Konsistenz oder *Intra*raterreliabilität genannt). Mehr Variation als erwartet kommt z.B. dann zustande, wenn ein ansonsten strenger Beurteiler bei einer geringen Anzahl von leistungsschwachen Pbn hohe Einstufungen vornimmt, also milde urteilt.

Tabelle 4.
Strenge und Konsistenz von 18 Beurteilern

Beurteiler	Strenge	Standard-fehler	Konsis-tenzindex A: Infit	Konsis-tenzindex B: Outfit	Korrigier-tes Ur-teilsmittel	Anzahl der Einzelur-teile
16	3.00	0.30	1.0	0.8	2.86	54
13	2.47	0.20	0.8	0.7	3.00	111
05	2.09	0.19	1.1	1.0	3.09	123
15	1.97	0.23	1.4	1.5	3.12	75
14	1.32	0.17	1.2	1.2	3.30	135
09	1.13	0.17	0.8	0.8	3.35	138
18	1.03	0.27	1.4	1.5	3.38	60
04	0.27	0.23	0.9	0.8	3.60	72
11	–0.07	0.26	0.8	0.8	3.70	57
08	–0.27	0.18	1.0	1.1	3.75	132
17	–0.78	0.17	0.8	0.8	3.88	138
10	–0.91	0.18	1.0	1.0	3.91	129
02	–1.28	0.24	1.2	1.1	4.00	75
12	–1.35	0.18	1.1	1.1	4.01	123
07	–1.62	0.15	1.0	0.9	4.08	204
03	–2.10	0.19	0.8	0.7	4.19	120
06	–2.12	0.26	0.7	0.7	4.20	63
01	–2.78	0.27	1.1	1.2	4.37	63

Werden die von Linacre (2002) vorgeschlagenen Grenzwerte für die Diagnose von Inkonsistenz im Urteilsverhalten zugrunde gelegt, so ist ersichtlich, dass sich die Beurteiler insgesamt durch ein hohes Maß an Konsistenz auszeichnen (d.h., kein einziger Beurteiler hat einen Infit- oder Outfit-Wert kleiner 0.5 bzw. größer 1.5). Selbst die Anwendung sehr restriktiver Grenzwerte (0.7/1.3) ändert an dieser Gesamteinschätzung kaum etwas. Lediglich bei zwei Fällen zeigen sich unter diesem engeren Blickwinkel Modellabweichungen: Beurteiler 15 und 18 zeigen größere Urteilsschwankungen, als nach dem Modell zu erwarten war; die Infit-Werte betragen jeweils 1.4 (40% mehr Variation als erwartet); die Outfit-Werte belaufen sich jeweils auf 1.5 (50% mehr Variation als erwartet). Daher ist bei diesen Beurteilern eine (relativ schwache) *Neigung zu Extremurteilen* zu vermuten. Beurteiler 13, 3 und 5 offenbaren dagegen tendenziell weniger Variation als erwartet (Infit- bzw. Outfit-Werte von 0.7, 30% weniger Variation als erwartet); hier wäre eher ein *Halo*- oder *Zentraleffekt*, d.h. eine Tendenz anzunehmen, bei der Beurteilung die mittleren TDN-Stufen zu bevorzugen.

Tabelle 4 liefert noch weitere Informationen. In Spalte 2 sind die ihrer Größe nach geordneten Logitwerte für die Strenge eines jeden Beurteilers angegeben (wie sie der Darstellung im Facettenraum zugrunde lagen). Mit jeder Schätzung eines Wertes des Strengeparameters ist ein Schätzfehler verbunden, der in standardisierter Form in Spalte 3 der Tabelle mitgeteilt wird. Wie allgemein in der statistischen Inferenz üblich, sinkt die Fehlergröße mit steigender Anzahl von Beobachtungen bzw. Einzelurteilen (letzte Spalte). In der vorletzten Spalte finden sich die korrigierten Urteilsmittel. Anhand dieser Mittelwerte erfährt die unterschiedliche Strenge der Beurteiler eine anschauliche Darstellung. Es handelt sich um den fairen Durchschnitt aller Einstufungen eines gegebenen Beurteilers, d.h. um die von einem Beurteiler durchschnittlich vorgenommene Einstufung unter Berücksichtigung der Fähigkeitsmaße der jeweils beurteilten Pbn. So zeigt etwa der Vergleich zwischen dem strengsten und dem mildesten Beurteiler, dass ihre Ratings im Durchschnitt um ca. eineinhalb Stufen auf der TDN-Skala differieren (4.37 – 2.86 = 1.51 TDN-Skalenpunkte).

5.3. Beurteilerstrenge und -konsistenz in anderen TestDaF-Prüfungen

Um abschätzen zu können, inwieweit die festgestellte Heterogenität der Beurteiler hinsichtlich ihrer Strenge wie auch die hohe Beurteilerkonsistenz spezifisch für die hier betrachtete TestDaF-Prüfung T002 sind, oder ob es sich dabei eher um ein typisches Ergebnismuster handelt, sind in Tabelle 5 die für eine Beantwortung dieser Frage relevanten Angaben zu fünf weiteren TestDaF-Prüfungen (T003 bis T007) zusammengestellt.

Tabelle 5.
Strenge und Konsistenz der Beurteiler in verschiedenen TestDaF-Prüfungen

Statistik	T002	T003	T004	T005	T006	T007
Anzahl der Beurteiler	18	29	28	32	21	21
Strenge (max/min)[a]	3.00/ –2.78	2.17/ –2.08	3.31/ –6.64	2.18/ –2.92	4.66/ –3.58	2.07/ –3.37
Homogenitätstest[b]	1118.5*	1828.5*	2371.6*	2052.9*	1682.3*	971.6*
Klassenseparation[c]	10.63	9.59	12.25	10.24	11.49	8.88
Separationsreliabilität[d]	0.98	0.98	0.99	0.98	0.99	0.98
Konsistenzindex A						
$0.7 \leq$ Infit ≤ 1.3	16	29	25	29	18	21
$0.5 \leq$ Infit ≤ 1.5	18	29	28	32	21	21
Konsistenzindex B						
$0.7 \leq$ Outfit ≤ 1.3	16	27	21	26	15	19
$0.5 \leq$ Outfit ≤ 1.5	18	29	27	32	19	21

Anmerkung. [a] Logitwerte. [b] Chi-Quadrat-Test (df = n – 1). [c] Anzahl statistisch reliabel unterscheidbarer Klassen von Beurteilern. [d] Genauigkeit, mit der die Strengewerte voneinander unterschieden werden können. * $p < .01$.

Die Prüfung T003 fand im April 2002 mit 1383 Teilnehmern statt; die Zahl der Beurteiler belief sich im Schriftlichen Ausdruck auf 29. Die entsprechenden Zahlen für die weiteren Prüfungen lauten: T004 (September 2002) mit 719 Teilnehmern und 28 Beurteilern, T005 (November 2002) mit 1095 Teilnehmern und 32 Beurteilern, T006 (Februar 2003) mit 1098 Teilnehmern und 21 Beurteilern sowie T007 (April 2003) mit 1554 Teilnehmern und 21 Beurteilern.

Der erste Vergleich bezieht sich auf die jeweiligen Maxima und Minima der Logitwerte für den Strengeparameter. Es ist leicht zu erkennen, dass die große Unterschiedlichkeit der Strengewerte in T002 (mit einer Streubreite von 5.78 Logits) ein ganz und gar typisches Ergebnis darstellt. Die Streubreite in den anderen Prüfungen bewegt sich zwischen 9.95 Logits (T004) und 4.25 Logits (T003). Der Homogenitätstest fällt bei allen Prüfungen hochsignifikant aus, d.h., die Annahme homogener Strengemaße ist zurückzuweisen (die Beurteiler unterscheiden sich überzufällig voneinander). Auch der Index der Klassenseparation zeichnet ein eindeutiges Bild: Die Zahl der statistisch reliabel unterscheidbaren Klassen von Beurteilern liegt bei rund 9 oder darüber. Wären die Beurteiler innerhalb einer Prüfung austauschbar, würden die Beurteiler also eine einzige, hinsichtlich ihrer Strengetendenzen homogene Gruppe bilden, dann sollte der Wert dieses Indexes nur unwesentlich mehr als 1 betragen. Die

Separationsreliabilität unterstreicht, dass die Beurteiler anhand ihrer Strengemaße sehr genau unterschieden werden können (maximal kann die Reliabilität den Wert 1 annehmen). Bei Gültigkeit der Homogenitätsannahme sollte diese Reliabilität gegen 0 gehen.

Auch im Hinblick auf die Stabilität der Beurteilerkonsistenz über die verschiedenen TestDaF-Prüfungen hinweg bestätigen sich die für T002 berichteten Ergebnisse. In keinem einzigen Fall liegen die Werte der Infit-Fehlerstatistik außerhalb der Grenzen des 0.5/1.5-Intervalls. Bei restriktiverer Definition der Intervallgrenzen gibt es nur in ganz wenigen Fällen Hinweise auf (leicht) inkonsistentes Urteilsverhalten. Kaum anders sehen die Ergebnisse für die Outfit-Fehlerstatistik aus. Lediglich bei Verwendung des engen Intervalls treten bei den Prüfungen T004 bis T006 etwas mehr Modellabweichungen auf. Dabei ist allerdings zu berücksichtigen, dass dieser Index gezielt das Vorkommen von sog. Ausreißern (d.h. von vereinzelt auftretenden starken Modellabweichungen in den extremen Urteilskategorien) erfassen soll. Üblicherweise kommt daher den Outfit-Werten auch deutlich weniger Gewicht bei der Klärung der Konsistenzfrage zu als den Werten des Infit-Index.

6. Korrekturverfahren im Vergleich

Die bislang berichteten Befunde belegen eindeutig die große Unterschiedlichkeit der Beurteiler in der Strenge bzw. Milde ihrer Leistungseinstufungen. Wie wirkt sich nun die Berücksichtigung bzw. Nichtberücksichtigung der nachgewiesenen Strengedifferenzen auf die Leistungsmessung entlang der TDN-Skala aus? Wie schneiden die Pbn im Vergleich unterschiedlicher Korrekturverfahren ab? Anhand zweier realer und zugleich typischer Fallbeispiele (TestDaF-Prüfung T002, Schriftlicher Ausdruck) soll im Folgenden der Einfluss des verwendeten Korrekturverfahrens auf das Prüfungsergebnis, also auf die in diesem Fertigkeitsbereich zugewiesene TDN-Stufe, aufgezeigt werden.

Die betrachteten Korrekturverfahren sind (a) das bei Sprachprüfungen weit verbreitete *Drittkorrekturverfahren* (im Folgenden auch „traditionelles Verfahren I“), (b) das *arithmetische Mittelungsverfahren* („traditionelles Verfahren II“) und (c) das aus dem MFRM-Modell abgeleitete *Multifacetten-Korrekturverfahren*. Drittkorrektur- und Mittelungsverfahren heißen „traditionell“, weil sie die Rohdaten, also die von den Beurteilern abgegebenen Bewertungen, *unmittelbar* für die TDN-Stufenzuweisung verwenden, d.h. in der Tradition der KTT stehend auf Parameterschätzung und nachfolgende Strengekorrektur komplett verzichten.

Im Drittkorrekturverfahren zählt als Gesamtergebnis bei übereinstimmenden Bewertungen die identisch vergebene TDN-Stufe; bei Nichtübereinstimmung

erfolgt eine Drittkorrektur zur endgültigen Festlegung der TDN-Stufe. Dieses Verfahren wurde in der Prüfung T002 angewendet. Das arithmetische Mittelungsverfahren verzichtet dagegen ganz auf die Durchführung einer Drittkorrektur bei nichtübereinstimmenden Bewertungen aus den ersten beiden Korrekturen. Statt dessen wird in jedem Fall das arithmetische Mittel über die sechs Einzelbewertungen (2 Korrekturen × 3 Kriterien) gebildet. Dieses Mittel wird auch „beobachteter Durchschnitt" genannt. Die Zuweisungsregel zur Bestimmung der TDN-Stufe lautet: „unter TDN 3" bei einem Mittelwert (M) kleiner 2.50, TDN 3 bei $2.50 \leq M \leq 3.49$, TDN 4 bei $3.50 \leq M \leq 4.49$ und TDN 5 bei einem M gleich oder größer 4.50.

Im Unterschied hierzu werden im Multifacetten-Korrekturverfahren die TDN-Bewertungen nach den einzelnen Kriterien gemäß des MFRM-Modells kalibriert, sodass Aussagen über die Strenge bzw. Milde der beteiligten Beurteiler getroffen und faire, strengekorrigierte Durchschnitte berechnet werden können. Unter Anwendung der oben definierten Zuweisungsregel für Mittelwerte wird die endgültige TDN-Stufe festgelegt.

Tabelle 6 zeigt den Vergleich zwischen den drei Korrekturverfahren für die Bewertung der Leistung von Pb 269.

Tabelle 6.
Korrekturverfahren im Vergleich: Fall A – Nichtübereinstimmung

Datenbasis			**Traditionelles Verfahren I**		**Trad. Verf. II**	**Multifacetten-Korrekturverfahren**	
Pb	Beurteiler	Kriterien GE–BdA–SR	Gesamt (TDN)	Drittkorr. (TDN)	Beob. Durchschnitt (TDN)	Strenge (Logits)	Fairer Durchschnitt (TDN)
269	03	4–5–4	TDN 4	**TDN 4** (4–3–4)	**TDN 5** (4.50)	–2.10 (mild)	**TDN 4** (4.07)
	12	5–5–4	TDN 5			–1.35 (mild)	

*Anmerkung***.** Die Daten stammen aus der TestDaF-Prüfung T002 (Schriftlicher Ausdruck; Pbn- und Korr.-Nr. geändert). Traditionelles Verfahren I = 2 Korrekturen plus 1 Drittkorrektur bei Nichtübereinstimmung. Traditionelles Verfahren II = 2 Korrekturen plus Ermittlung des beobachteten Durchschnitts. Multifacetten-Korrekturverfahren = 2 Korrekturen plus Ermittlung des fairen Durchschnitts (gemäß MFRM-Modell). GE = Gesamteindruck. BdA = Be-

handlung der Aufgabe. SR = Sprachliche Realisierung. TDN = TestDaF-Niveaustufe.

Die Leistung dieses Pb im schriftlichen Ausdruck wurde von den Beurteilern 03 und 12 bewertet. Da die Gesamtbewertungen um 1 TDN-Stufe differierten, wurde gemäß des traditionellen Verfahrens I eine dritte Korrektur vorgenommen; diese führte zu dem Endergebnis TDN 4. Der beobachtete Durchschnitt (traditionelles Verfahren II) belief sich auf 4.50, sodass nach der obigen Zuweisungsregel die TDN-Stufe 5 zu vergeben wäre.

Wie würde das Ergebnis für Pb 269 bei Anwendung des Multifacetten-Korrekturverfahrens aussehen? Die Strengeanalyse ergab für Beurteiler 03 einen Logitwert von –2.10, für Beurteiler 12 resultierten –1.35 Logits. Damit waren beide als eher *mild* einzustufen. Mit anderen Worten, die Bewertungen beider Beurteiler tendierten zu einer *Überschätzung* der Leistungsfähigkeit. Diese Urteilstendenzen wurden bei der Berechnung des fairen Durchschnitts berücksichtigt. Dieser belief sich auf 4.07, sodass (nach derselben Zuweisungsregel wie beim beobachteten Durchschnitt) die TDN-Stufe 4 resultierte.

Einen anderen Fall gibt Tabelle 7 wieder.

Tabelle 7.
Korrekturverfahren im Vergleich: Fall B – Übereinstimmung

Datenbasis			Traditionelles Verfahren I		Trad. Verf. II	Multifacetten-Korrekturverfahren	
Pb	Erst- u. Zweitkorr.	Kriterien GE–BdA–SR	Gesamt (TDN)	Keine Drittkorr. (TDN)	Beob. Durchschnitt (TDN)	Strenge (Logits)	Fairer Durchschnitt (TDN)
034	13	4–4–4	TDN 4	**TDN 4**	**TDN 4** (3.83)	2.47 (streng)	**TDN 5** (4.52)
	16	4–4–3	TDN 4			3.00 (streng)	

Anmerkung. Die Daten stammen aus der TestDaF-Prüfung T002 (Schriftlicher Ausdruck; Pbn- und Korr.-Nr. geändert). Traditionelles Verfahren I = 2 Korrekturen plus 1 Drittkorrektur bei Nichtübereinstimmung. Traditionelles Verfahren II = 2 Korrekturen plus Ermittlung des beobachteten Durchschnitts. Multifacetten-Korrekturverfahren = 2 Korrekturen plus Ermittlung des fairen Durchschnitts (gemäß MFRM-Modell). GE = Gesamteindruck. BdA = Be-

handlung der Aufgabe. SR = Sprachliche Realisierung. TDN = TestDaF-Niveaustufe.

Die Leistung von Pb 034 wurde von den Beurteilern 13 und 16 in übereinstimmender Weise bewertet. Beide kamen zum Ergebnis, dass eine Einstufung nach TDN 4 angemessen sei. Eine Drittkorrektur war daher überflüssig. Auch das arithmetische Mittelungsverfahren führte mit einem beobachteten Durchschnitt von 3.83 zur Endbewertung TDN 4.

Die Prüfungsleistung von Pb 034 wurde also nach beiden traditionellen Korrekturverfahren *identisch* eingestuft. Bei soviel Übereinstimmung sollte an dieser Bewertung nichts auszusetzen sein. Doch weit gefehlt: Hinter der nahezu perfekten Übereinstimmung zwischen Beurteilern 13 und 16 verbirgt sich ein hohes Maß an *Ungerechtigkeit* in der Bewertung der Prüfungsleistung. Wie die Strengewerte der Beurteiler zeigen, zeichneten sich beide durch eine ausgeprägte Tendenz zur Strenge aus. Der Schreibbogen von Pb 034 wurde von den in dieser Prüfung *strengsten* Beurteilern bewertet (vgl. Abb. 2 und Tab. 4). Der faire Durchschnitt von 4.52 korrigiert die klare Unterschätzung der Fähigkeit dieses Pb. Abweichend von den Ergebnissen der beiden traditionellen Korrekturverfahren würde das Multifacetten-Korrekturverfahren eine Zuordnung zur höchsten TDN-Stufe sicherstellen.

7. Voraussetzungen des Multifacetten-Korrekturverfahrens

7.1. Konsistente Beurteilungen

Eine Voraussetzung für die Ermittlung von Maßen der Beurteilerstrenge und die entsprechende Korrektur der abgegebenen Einstufungen im Rahmen eines Multifacetten-Korrekturverfahrens ist das Vorliegen hinreichend hoher Konsistenz der Bewertungen. Erst wenn die Konsistenzbedingung als erfüllt betrachtet werden kann (hierzu sind, wie gezeigt wurde, die Infit- und Outfit-Statistiken heranzuziehen), lassen sich die Strengemaße sinnvoll interpretieren und bei der abschließenden Leistungseinstufung berücksichtigen.

Einer der Vorzüge einer Multifacetten-Rasch-Analyse besteht darin, detaillierte *Rückmeldung* an die Beurteiler geben zu können. Die Rückmeldung hat das Ziel, Beurteilern im Anschluss an eine TestDaF-Prüfung Informationen über wesentliche Aspekte ihres Korrektur- oder Bewertungsverhaltens an die Hand zu geben. Diese Informationen können ihnen helfen, Fragen bezüglich der eigenen Bewertungsleistung zu beantworten, möglicherweise vorhandene Tendenzen bei ihren Bewertungen aufzuzeigen und damit auch die Qualität des gesamten Korrekturprozesses zu erhöhen bzw. auf Dauer zu sichern. Beurteiler erhalten darüber Auskunft, wie streng bzw. milde, aber auch wie konsi-

stent bzw. inkonsistent ihre Bewertungen der Prüfungsleistungen ausgefallen sind.

Wie bereits ausgeführt, hat sich der Strengeeffekt in der Forschung zu Sprachprüfungen als ein sehr starker und robuster Effekt erwiesen. Selbst zeitlich ausgedehnte, intensive Schulungen oder Trainings können diesen Effekt nur geringfügig, wenn überhaupt, mindern. Ziel von Beurteilertrainings sollte es daher *nicht* sein, das Korrekturverhalten dahingehend zu verändern, dass *alle* Beurteiler *denselben* Bewertungsstandard verwenden (Stahl/Lunz, 1996). Eine der vorrangigen Aufgaben solcher Trainings sollte es sein, die Bedeutung und Anwendung der jeweiligen Beurteilungskriterien (Unterkriterien und übergeordnete Kriterien) zu klären und den korrekten Gebrauch der Ratingskala zu erläutern. Die Aufmerksamkeit sollte sich also auf die Erhöhung bzw. Stabilisierung der Konsistenz *innerhalb* der Beurteiler und nicht auf die Beseitigung der Strengeunterschiede *zwischen* den Beurteilern richten.

7.2. Vergleichbarkeit der Beurteiler

Neben der Konsistenz der Bewertungen muss *Vergleichbarkeit* der Beurteiler gegeben sein. Technisch gesehen handelt es sich um die Voraussetzung der *Konnektivität* der zugrunde liegenden Datenmatrix (Linacre/Wright, 2002). Eine Datenmatrix wird konnektiv genannt, wenn (anschaulich gesprochen) ein Netzwerk von Verbindungen zwischen Elementen von Facetten besteht, das eng genug geknüpft ist, um alle Beurteiler über gemeinsame Kriterien (Items, Aufgaben) und gemeinsame Pbn direkt oder indirekt miteinander zu verbinden (Engelhard, 1997; Lunz/Wright/Linacre, 1990; vgl. auch Wright/Stone, 1979: 98ff).

Nur wenn diese Eigenschaft vorliegt, ist eine eindeutige Interpretation der Ergebnisse möglich, d.h., nur im Falle einer konnektiven Datenmatrix lassen sich alle Elemente aller Facetten in einem gemeinsamen Bezugsrahmen darstellen. Mangelnde Konnektivität hätte zur Folge, dass einzelne Elemente in voneinander separierte Teilmengen zusammengefasst würden und Vergleiche zwischen Elementen nur innerhalb, nicht aber zwischen diesen Teilmengen sinnvoll durchführbar wären. So bliebe z.B. unklar, ob ein Beurteiler deshalb höhere Einstufungen vorgenommen hat, weil er oder sie strenger als andere Beurteiler ist, oder ob die beurteilten Pbn einer stärkeren Leistungsgruppe angehörten.

Um eine konnektive Datenmatrix herzustellen, ist es in der Regel ausreichend, alle Beurteiler eine kleine Anzahl von Pbn-Leistungen beurteilen zu lassen (Myford/Wolfe, 2000). Diese gemeinsamen Beurteilungen (auch „Vergleichskorrekturen“ genannt) stellen die nötige Verbindung zwischen den Elementen

der verschiedenen Facetten sicher. Unter Beachtung der Konnektivitätsbedingung lassen sich bei geeignet konstruierten Korrektur- oder Bewertungsplänen Einsparungen an Zeit und Kosten erreichen, die insbesondere für breit angelegte Sprach- und Leistungsprüfungen von erheblicher Bedeutung sein können (Linacre/Wright, 2002).

8. Schlussbemerkungen

Im Rahmen von Sprachprüfungen zielen Leistungsbeurteilungen darauf ab, möglichst genaue und verlässliche Aussagen über die sprachliche Leistungsfähigkeit zu erlauben. Merkmale der Testsituation, die nichts mit der zu messenden Fähigkeit an sich zu tun haben, sollten nur einen vernachlässigbar geringen Einfluss auf die Beurteilungen nehmen. Damit also das Leistungsergebnis eines bestimmten Pb in der intendierten Weise interpretiert werden kann, ist sicherzustellen, dass der ermittelte Fähigkeitswert soweit wie möglich unabhängig ist von der Wirkung leistungsirrelevanter Faktoren.

Ganz allgemein gesprochen ist für die psychometrische Qualität von Leistungsbeurteilungen entscheidend, wie gut es gelingt, die verschiedenen Quellen der Variabilität in ihrem Einfluss auf die Beurteilungen zu identifizieren und zu kontrollieren. Die Multifacetten-Rasch-Analyse (Linacre, 1989; Linacre/Wright, 2002) stellt das methodische Instrumentarium bereit, das für eine differenzierte Qualitätskontrolle erforderlich ist. Mit dem Korrekturverfahren, das aus diesem Modellansatz abgeleitet ist, verbinden sich nicht nur die Vorteile einer fairen Leistungsbeurteilung und detaillierten Rückmeldung an die Beurteiler, sondern auch die Vorteile einer breit angelegten, alle relevanten Facetten der Prüfungssituation umfassenden Sicherung und Kontrolle der Qualität von Leistungsbeurteilungen.

Stichwortartig seien die wesentlichen Vorzüge einer Multifacetten-Rasch-Analyse im Kontext von Sprachprüfungen zusammengefasst.

(a) Messung der Beurteilerstrenge, der Personenfähigkeit und der Kriterien- bzw. Aufgabenschwierigkeit auf einer gemeinsamen linearen Skala (Logitskala).
(b) Überprüfung der Annahme einer strengehomogenen Beurteilergruppe.
(c) Erfassung der Konsistenz des Bewertungsverhaltens jedes einzelnen Beurteilers.
(d) Faire Messung der Leistungsfähigkeit der Pbn durch Berücksichtigung der Beurteilerstrenge.
(e) Rückmeldung der Ergebnisse zu den Strenge- und Konsistenzanalysen an die Beurteiler.

(f) Optimierung von Beurteilertrainings durch Fokussierung auf die Urteilskonsistenz.
(g) Kosten- und Zeitersparnis durch Anwendung geeigneter Korrekturpläne.

Perspektivisch ist der Blick auf die Entwicklung möglichst objektiver, genauer und fairer Systeme zur differenzierten Leistungsbeurteilung zu richten. Beurteilungen sind ein fester Bestandteil von Konzepten der Leistungsevaluation; es ist unschwer abzusehen, dass der Stellenwert von Beurteilungen im Rahmen solcher Evaluationen in den nächsten Jahren noch weiter steigen wird. Umso mehr wird es darauf ankommen, Einflüsse subjektiver Urteilstendenzen soweit wie möglich zurückzudrängen. Nur eine wissenschaftlich fundierte Leistungsmessung, wie sie das Multifacetten-Rasch-Modell ermöglicht, wird die erforderliche Kontrolle und Sicherung der Qualität von Beurteilungen auf Dauer gewährleisten können (Eckes, 2003; Engelhard, 2002; Paulukonis/Myford/Heller, 2000). Die fortlaufende Qualitätssicherung von Leistungsbeurteilungen ist dabei nicht nur für Sprachprüfungen zu fordern, sondern auch für all jene Situationen, in denen die Einschätzung der individuellen Leistungsfähigkeit durch Beurteiler Routine ist und weitreichende Konsequenzen für die Beurteilten hat (z.B. in Schulen, Universitäten, Betrieben).

Eine andere Perspektive betrifft die Notwendigkeit, die angewandte Forschung im Bereich der Leistungsbeurteilung zu intensivieren. Es ist für empirische Forschung geradezu typisch, dass sie insbesondere in einem frühen Stadium mehr Fragen aufwirft, als sie Antworten zu geben in der Lage ist. Ein im Kontext der Multifacetten-Rasch-Analyse noch relativ wenig erforschtes Gebiet betrifft die Beurteilerstrenge. So wurde in dieser Arbeit zwar eine statistisch-operationale Definition von Strenge bzw. Milde gegeben, aber eine inhaltliche Betrachtung dieses zentralen Konstrukts erfolgte nicht. Unter Inhaltsaspekten fassten Stahl/Lunz (1996) Beurteilerstrenge als das Gesamt der individuellen, sozialen, lern- und erfahrungsabhängigen Faktoren, die auf das Muster des Urteilsverhaltens Einfluss nehmen. Eine derart breite Konzeption verlangt nach Präzision; vor allem ist zu klären, welche der postulierten Faktoren in welchem Maß den individuellen Grad der Strenge bestimmen (Brown, 1995). Damit verbindet sich natürlich auch die Frage nach der zeitlichen und transsituationalen Stabilität der Beurteilerstrenge (Lumley/McNamara, 1995; O'Neill/Lunz, 2000; Wilson/Case, 2000). Möglicherweise ist dem Konstrukt einer *differentiellen Strenge* der Vorzug vor einer kontextinvarianten Konzeption zu geben. Hier könnte die Differentielle Psychologie und Persönlichkeitsforschung (vgl. z.B. Amelang/Bartussek, 2001) wichtige Anregungen geben.
Schließlich ist zu beachten, dass die Tendenz zur Strenge bzw. Milde nicht die einzige Urteilstendenz ist, die mangelnde Übereinstimmung nach sich ziehen

kann. Andere, oft beschriebene (und hier nur kurz wiederzugebende) Urteilstendenzen sind die Zentraltendenz, die Extremtendenz und der Halo-Effekt. Die *Zentraltendenz* (oder *Tendenz zur Mitte*) betrifft die Neigung, die mittleren Kategorien einer mehrstufigen Ratingskala bevorzugt zu verwenden bzw. die extremen Kategorien zu vermeiden. Umgekehrt wird mit der *Extremtendenz* die Neigung beschrieben, gehäuft extreme Urteilskategorien zu verwenden. Unter einem *Halo-Effekt* ist die Tendenz zu verstehen, Einstufungen auf unterschiedlichen Merkmalen von einem ganz bestimmten Urteil (z.B. von einer positiven oder negativen Gesamtbewertung einer Person oder eines hervorstechenden Merkmals der Person) leiten zu lassen. In der Literatur finden sich zwar Vorschläge, auch diese Urteilstendenzen im Rahmen von Multifacetten-Rasch-Analysen zu identifizieren bzw. zu kontrollieren (Engelhard, 1994; Wolfe/Chiu/Myford, 2000), doch die Forschung hierzu steckt noch in den Anfängen.

9. Literatur

Amelang, N./Bartussek, D. (2001). *Differentielle Psychologie und Persönlichkeitsforschung* (5. Aufl.). Stuttgart: Kohlhammer.

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

Artelt, C./Stanat, P./Schneider, W./Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert/E. Klieme/M. Neubrand/M. Prenzel/U. Schiefele/W. Schneider/P. Stanat/K.-J. Tillmann/M. Weiß (Hrsg.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69–137). Opladen: Leske + Budrich.

Baumert, J./Klieme, E./Neubrand, M./Prenzel, M./Schiefele, U./Schneider, W./Stanat, P./Tillmann, K.-J./Weiß, M. (Hrsg.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Birkel, P./Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, *49*, 219–224.

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

Bortz, J./Döring, N. (2002). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (3. Aufl.). Berlin: Springer-Verlag.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*, 1–15.

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

Eckes, T. (2003). Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse. *Fremdsprachen und Hochschule, 69,* 43-68.

Eckes, T. (2004). Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im „Test Deutsch als Fremdsprache“ (TestDaF). *Diagnostica*, *50*, 65-77.

Eckes, T. (in Druck). Rasch-Modelle zur C-Test-Skalierung. In R. Grotjahn (Hrsg.), *The C-test: Theory, empirical research, applications.* Frankfurt: Lang

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

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

Engelhard, G. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, *1*, 19–33.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal/T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Erlbaum.

Europarat. (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: Lernen, lehren, beurteilen*. Berlin: Langenscheidt.

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen*. Bern: Huber.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Gulliksen, H.. (1950). *Theory of mental tests*. New York: Wiley.

Hambleton, R K./Robin, F./Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In H. E. A. Tinsley/S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 553–581). San Diego, CA: Academic Press.

Hedges, L. V./Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Hoyt, W. T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods*, *5*, 64–86.

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

Ingenkamp, K. (1989). Die diagnostische Problematik des Aufsatzes als Prüfungsinstrument und die Bemühungen zur Verbesserung der Auswertungsqualität. In K. Ingenkamp, *Diagnostik in der Schule: Beiträge zu Schlüsselfragen der Schülerbeurteilung* (S. 127–149). Weinheim: Beltz.

Ingenkamp, K. (Hrsg.). (1995). *Die Fragwürdigkeit der Zensurengebung: Texte und Untersuchungsberichte* (9. Aufl.). Weinheim: Beltz.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342.

Khattri, N./Sweet, D. (1996). Assessment reform: Promises and challenges. In M. B. Kane/R. Mitchell (Eds.), *Implementing performance assessment: Promises, problems, and challenges* (pp. 1–22). Mahwah, NJ: Erlbaum.

Lehmann, R. H. (1988). Zuverlässigkeit und Generalisierbarkeit von Aufsatzbewertungen. *Empirische Pädagogik*, *2*, 349–365.

Lehmann, R. H. (1990). Aufsatzbeurteilung – Forschungsstand und empirische Daten. In K. Ingenkamp/R. S. Jäger (Hrsg.), *Tests und Trends: Jahrbuch der Pädagogischen Diagnostik* (Bd. 8, S. 64–94). Weinheim: Beltz.

Lienert, G. A./Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Psychologie Verlags Union.

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

Linacre, J. M. (1999). *A user's guide to Facets: Rasch measurement computer program*. Chicago: MESA Press.

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

Linacre, J. M. (2003). Size vs. significance: Standardized chi-square fit statistic. *Rasch Measurement Transactions*, *17*(1), 918.

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

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

Lunz, M. E./Wright, B. D./Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*, 331–345.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

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

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

Myford, C. M./Wolfe, E. W. (2000). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs* (TOEFL Technical Report, TR-15). Princeton, NJ: Educational Testing Service.

O'Neill, T. R./Lunz, M. E. (2000). A method to study rater severity across several administrations. In M. Wilson/G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 135–146). Stamford, CT: Ablex.

Paulukonis, S. T./Myford, C. M./Heller, J. I. (2000). Formative evaluation of a performance assessment scoring system. In M. Wilson/G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 15–40). Stamford, CT: Ablex.

Quetz, J. (2003). A1 – A2 – B1 – B2 – C1 – C2: Der Gemeinsame europäische Referenzrahmen. *Deutsch als Fremdsprache*, *40*, 42–48.

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

Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion (2. Aufl.)*. Bern: Huber.

Saal, F. E./Downey, R. G./Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.

Stahl, J. A./Lunz, M. E. (1996). Judge performance reports: Media and message. In G. Engelhard/M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 113–125). Norwood, NJ: Ablex.

Steyer, R./Eid, M. (2001). *Messen und Testen* (2. Aufl.). Berlin: Springer-Verlag.

Weiss, R. (1965). Über die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen. *Schule und Psychologie*, *18*, 257–269.

Wilson, M./Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson/G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 113–133). Stamford, CT: Ablex.

Wirtz, M./Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.

Wolfe, E. W./Chiu, C. W. T./Myford, C. M. (2000). Detecting rater effects in simulated data with a multifaceted Rasch rating scale model. In M. Wilson/G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 147–164). Stamford, CT: Ablex.

Wright, B. D./Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B. D./Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D./Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.