Unser Themenschwerpunkt „Testen und Prüfen im Bereich Deutsch als Fremd- und Zweitsprache“, der im Heft 2/2010 eröffnet wurde, wird in diesem Heft durch die Beiträge von Thomas Eckes und Karin Vogt sowie durch die Rezension von Katrin Wisniewski abgeschlossen.

Die Redaktion

Thomas Eckes

Facetten der Genauigkeit. Zur Reliabilität der Beurteilung fremdsprachlicher Leistungen

Die Beurteilung fremdsprachlicher Leistungen unterliegt einer Reihe von Fehlern. Diese Fehler äußern sich darin, dass selbst erfahrene und geschulte Beurteiler bei ein und derselben Leistung zu unterschiedlichen Bewertungen kommen. Im vorliegenden Beitrag werden traditionelle Ansätze zur Lösung des Problems, insbesondere Beurteilerschulung, wiederholte Beurteilung und Berechnung eines Maßes der Reliabilität, als unzureichend kritisiert und neuere Methoden zur Kontrolle bzw. Korrektur von Urteilsfehlern besprochen.

Ratings of language performance are subject to a number of errors. These errors can be seen from the fact that even experienced and trained raters come up with different ratings for the very same performance. In this paper, it is argued that traditional approaches to solving the problem like rater training, repeated ratings, and computing an index of interrater reliability are insufficient. More recent methods are suggested that can be used to control or compensate for judgemental errors.

1 Einleitung

Wenn fremdsprachliche Leistungen, z. B. in Form eines Aufsatzes oder in der mündlichen Kommunikation, zu beurteilen sind, stellt sich unmittelbar die Frage nach der Reliabilität der Beurteilung. Würde ein Teilnehmer, dessen Leistung zwei oder mehr Beurteiler unabhängig voneinander einstufen, von allen Beurteilern dieselbe Bewertung erfahren? Oder würden die Bewertungen unterschiedlich ausfallen? Wie ließe sich das Ausmaß möglicher Unterschiede in den Bewertungen ermitteln? Die vorliegende Arbeit behandelt diese Fragen.[1] Allgemeiner gesprochen geht es um die Bestimmung der Reliabilität im Rahmen weitgehend standardisierter Verfahren der Leistungsbeurteilung. Für solche Verfahren sind auch die Bezeichnungen „beurteilergestützte Leistungsmessung“ (vgl. Engelhard 2002; Eckes 2009; 2010) oder „Performanztest“ (vgl. McNamara 1996; Grotjahn 2000; Wigglesworth 2008) gebräuchlich.

Ein Grundproblem der beurteilergestützten Leistungsmessung liegt in der Subjektivität des Bewertungsprozesses und der damit verbundenen Anfälligkeit für verschiedene Arten von Urteilsfehlern. Urteilsfehler wie die Tendenz einzelner Beurteiler zu Strenge oder Milde können die Qualität eines Performanztests unter Umständen stark gefährden und die Aussagekraft der Testergebnisse mindern. Es ist daher unerlässlich, eine ausreichend hohe Beurteilerreliabilität nachzuweisen.[2] Entsprechende international verbindliche Qualitätsstandards sind seit langem verfügbar (vgl. z. B. Grotjahn 2000; 2007; Moosbrugger/Höfling 2007). Vor dem Hintergrund dieser Standards werden in diesem Beitrag verschiedene Methoden zur Bestimmung der Beurteilerreliabilität besprochen und Probleme ihrer Anwendung diskutiert.

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

[2] Beurteilerreliabilität wird auch Interraterreliabilität genannt, und zwar insbesondere dann, wenn diese Form der Reliabilität von der Reliabilität der Urteile eines bestimmten Beurteilers zu verschiedenen Zeitpunkten oder über verschiedene Teilnehmer hinweg unterschieden werden soll; letztere heißt Intraraterreliabilität (oder Beurteilerkonsistenz).

2 *Methoden der Reliabilitätsbestimmung*

2.1 *Der traditionelle Ansatz*

Die übliche Antwort auf das notorische Problem der Urteilsfehler sieht drei methodische Schritte vor: Beurteilerschulung, wiederholte Leistungsbewertung durch unabhängige Beurteiler und Nachweis der Beurteilerreliabilität. Liegt hohe Reliabilität vor, dann – so lautet die weit verbreitete Annahme – herrscht Konsens zwischen den Beurteilern hinsichtlich der fremdsprachlichen Leistung und den sie auszeichnenden Merkmalen. Mit anderen Worten: Hohe empirisch nachgewiesene Reliabilität impliziert ein hohes Maß an Beurteilerübereinstimmung, woraus sich wiederum Genauigkeit der Urteile folgern ließe. Im Idealfall wären die Beurteiler untereinander austauschbar, d. h., es sollte keinen Unterschied machen, welche Beurteiler aus einer Gruppe gleichermaßen geschulter Beurteiler eine gegebene Leistung bewerteten (vgl. z. B. Wirtz/Caspar 2002).

Diese Logik hat jedoch gleich mehrere Schwächen. Erstens kann eine Erhöhung der Reliabilität von Beurteilungen zu dem unerwünschten Ergebnis führen, dass ihre Validität sinkt. Dies wäre z. B. der Fall, wenn sich Beurteiler bei einer Sprachprüfung im Hochschulkontext mehr oder weniger stillschweigend darauf verständigten, ihre Aufmerksamkeit stärker auf relativ einfach zu bewertende, aber weniger wichtige Leistungsmerkmale wie grammatikalische oder orthografische Korrektheit zu richten und wichtigere Merkmale, die eher mit unterschiedlichen Bewertungen einhergehen, wie Breite der sprachlichen Mittel oder Vollständigkeit der Aufgabenbearbeitung, zu vernachlässigen (vgl. z. B. Hamp-Lyons 2007; McNamara 1996).

Zweitens ist der Schluss von hoher Beurteilerübereinstimmung auf hohe Urteilsgenauigkeit selber alles andere als zwingend. So kann ein hohes Maß an Beurteilerübereinstimmung auch dadurch zustande kommen, dass Beurteiler gleich gerichteten Urteilsfehlern unterliegen. Die Beurteiler könnten z. B. ähnlich streng oder ähnlich milde urteilen. Hierauf wird weiter unten ausführlicher eingegangen.

Drittens gibt es nicht einige wenige, allgemein akzeptierte Methoden, um die Beurteilerreliabilität zu ermitteln, sondern eine Vielzahl von Methoden mit z. T. höchst unterschiedlichen statistischen Eigenschaften (vgl. z. B. Zegers 1991; Wirtz/Caspar 2002; Shoukri 2004; von Eye / Mun 2005; Wirtz 2006; Bramley 2007; Hayes/Krippendorff 2007; LeBreton/Senter 2008; Uebersax 2008). Diese unterschiedlichen Eigenschaften können dazu führen, dass zwei Methoden bei ein und demselben Datensatz gegenläufige Ergebnisse erbringen; d. h., die eine Methode könnte Ergebnisse liefern, die auf hohe Reliabilität verweisen, die andere Methode dagegen niedrige Reliabilität indizieren.

2.2 *Konsens- und Konsistenzmethoden*

Die kaum noch überschaubare Fülle an Methoden zur Bestimmung der Beurteilerreliabilität macht es zu keiner leichten Aufgabe, die für eine konkrete Untersuchung geeignete(n) Methode(n) zu wählen. Zur besseren Orientierung erscheint es daher sinnvoll, mit Tinsley/Weiss (1975; 2000) grob zwischen zwei Klassen von Methoden zu unterscheiden: Konsensmethoden und Konsistenzmethoden (vgl. auch Stemler/Tsai 2008).

Konsensmethoden erfassen das Ausmaß, in dem unabhängige Beurteiler gleiche Beurteilungen eines Teilnehmers oder einer Leistung abgeben; Konsensmaße der Beurteilerreliabilität geben den Grad der absoluten Korrespondenz der Urteile wieder. Im Unterschied hierzu erfassen Konsistenzmethoden das Ausmaß, in dem die beurteilten Teilnehmer oder Leistungen in der gleichen Relation zueinander stehen; Konsistenzmaße der Beurteilerreliabilität geben den Grad der relativen Korrespondenz der Urteile wieder.

Auch wenn Methoden beider Klassen in der Literatur nicht selten gleichbedeutend verwendet werden, können doch Konsens- und Konsistenzmaße stark differierende, bisweilen gegensätzliche Schlüsse über die Höhe der Reliabilität nahelegen. Z. B. könnte ein Beurteiler ein und dieselben Teilnehmerleistungen durchweg um ein oder zwei Skalenpunkte höher bewerten als ein anderer Beurteiler. Die relative Korrespondenz beider Reihen von Bewertungen wäre sehr hoch; Konsistenzmaße nähmen daher hohe Werte an. Übereinstimmung hätten die Beurteiler aber in keinem einzigen Fall erzielt, was in sehr niedrigen Werten von Konsensmaßen zum Ausdruck käme.

Tab. 1 gibt eine Übersicht über Konsens- und Konsistenzmaße, die teils im Zusammenhang mit fremdsprachlichen Performanztests,

Maß (Kürzel, Symbol)	Bezeichnung	Autor(en)	Wertebereich	Kurzbeschreibung	Weiterführende Literatur
				Konsensmaße	
P_o	Exakte Übereinstimmung	Vgl. z. B. Wirtz/Caspar (2002)	$0 \leq P_o \leq 1$	Anteil exakt übereinstimmender Urteile. Sehr einfach zu berechnen, anschaulich. Keine Korrektur für Zufallsübereinstimmung. Unterscheidet nur zwischen übereinstimmenden und nichtübereinstimmenden Urteilen. Verlangt lediglich nominalskalierte Urteile.	Fleiss/Levin/Paik (2003)
κ	Cohens Kappa	Cohen (1960)	$-1 \leq \kappa \leq 1$	Anteil exakt übereinstimmender Urteile nach Zufallskorrektur. Negative Werte besagen, dass zwei Beurteiler weniger übereinstimmen, als nach Zufall zu erwarten wäre. Extremwerte (–1, 1) werden nur erreicht, wenn die Randsummenverteilungen für beide Beurteiler identisch sind. Verlangt lediglich nominalskalierte Urteile.	Wirtz/Caspar (2002); Fleiss/Levin/Paik (2003); Mun (2005)
κ_w	Cohens gewichtetes Kappa	Cohen (1968)	$-1 \leq \kappa_w \leq 1$	Verallgemeinert Kappa, um nichtübereinstimmende Urteile (insbesondere bei Ratingskalen) abgestuft zu berücksichtigen. Nichtübereinstimmungen werden umso geringer gewichtet, je weiter die Urteile auseinanderliegen (Gewichte zwischen 0 und 1). Verlangt mindestens ordinalskalierte Urteile.	Wirtz/Caspar (2002); Fleiss/Levin/Paik (2003); Mun (2005)
RAI	Rater Agreement Index	Burry-Stock/Shaw/Laurie/Chissom (1996)	$0 \leq RAI \leq 1$	Primär konzipiert für Übereinstimmung im Falle von Ratingskalen. Basiert auf der Summe der absoluten Differenzen zwischen Skalenkategorien. Einfach zu berechnen. Geeignet für zwei oder mehr Beurteiler, ein oder mehr Kriterien und ein oder mehr Beurteilungsobjekte. Verlangt mindestens intervallskalierte Urteile.	Zegers (1991)
OAI	Overall (Raw) Agreement Index	Uebersax (2008)	$0 \leq OAI \leq 1$	Übereinstimmung bei dichotomen und polytomen Urteilen. Relativiert beobachtete Übereinstimmung an theoretisch möglicher Übereinstimmung. Einfach zu berechnen. Geeignet für zwei oder mehr Beurteiler und größere Anzahl von Beurteilungsobjekten. Verlangt lediglich nominalskalierte Urteile.	Jones (2005)
r_{WG}	Within-Group Agreement Index	James/Demaree/Wolf (1984)	$0 \leq r_{WG} \leq 1$	Übereinstimmung als proportionale Reduktion der Fehlervarianz. Relativiert beobachtete Varianz an Urteilsvarianz im Falle maximaler Nichtübereinstimmung. Anzahl der Beurteiler sollte mindestens 10 betragen. Geeignet für ein oder mehr Kriterien und ein Beurteilungsobjekt. Verlangt mindestens intervallskalierte Urteile.	Cohen/Doveh/Nahum-Shani (2009); Lüdtke/Robitzsch (2009)
AD_M	Average Deviation Index	Burke/Finkelstein/Dusig (1999)	$0 \leq AD_M$	Basiert auf der absoluten Differenz zwischen Einzelurteilen und dem Urteilsmittel in einer Beurteilergruppe. Kann in der Metrik der Ratingskala interpretiert werden. Höhere Werte zeigen niedrigere Übereinstimmung an. Geeignet für ein oder mehr Kriterien und ein Beurteilungsobjekt. Verlangt mindestens intervallskalierte Urteile.	Dunlap/Burke/Smith-Crowe (2003); Cohen/Doveh/Nahum-Shani (2009)

Maß (Kürzel, Symbol)	Bezeichnung	Autor(en)	Wertebereich	Kurzbeschreibung	Weiterführende Literatur
				Konsistenzmaße	
r	Produkt-Moment-Korrelation (Pearson-Korrelation)	Vgl. z. B. Wirtz/Caspar (2002)	$-1 \leq r \leq 1$	Gibt die Enge des linearen Zusammenhangs zwischen den Urteilen zweier Beurteiler an. Die Urteile können auf Skalen mit unterschiedlicher Metrik basieren (Interklassenkorrelation). Negative Werte verweisen auf tendenziell gegenläufige Einstufungen (Reliabilität wird gleich 0 gesetzt). Verlangt mindestens intervallskalierte Urteile.	Murphy/DeShon (2000); Wirtz/Caspar (2002)
τ_b	Kendalls Rangkorrelation (Tau-b)	Kendall (1948)	$-1 \leq \tau_b \leq 1$	Gibt die Enge des ordinalen Zusammenhangs zwischen den Urteilen zweier Beurteiler an (Übereinstimmung zwischen den Rangreihen). Berücksichtigt Rangbindungen. Negative Werte verweisen auf tendenziell gegenläufige Einstufungen (Reliabilität wird gleich 0 gesetzt). Verlangt mindestens ordinalskalierte Urteile.	Bortz/Lienert/Boehnke (2000); Wirtz/Caspar (2002)
W	Kendalls Konkordanzkoeffizient	Kendall (1948)	$0 \leq W \leq 1$	Gibt die Enge des ordinalen Zusammenhangs zwischen den Urteilen von mehr als zwei Beurteilern an (Übereinstimmung zwischen den Rangreihen innerhalb einer Beurteilergruppe). Berücksichtigt Rangbindungen. Kann auch für unvollständige Beurteilungspläne berechnet werden. Verlangt mindestens ordinalskalierte Urteile.	Bortz/Lienert/Boehnke (2000); von Eye / Mun (2005)
α	Cronbachs Alpha	Cronbach (1951)	$0 \leq \alpha \leq 1$	Schätzt den Anteil der beobachteten Varianz (d. h. Varianz der Urteile von zwei oder mehr Beurteilern), die auf systematische Unterschiede zwischen den Beurteilungsobjekten zurückgeht. Beurteiler werden wie Items eines objektiven Tests behandelt (interne Konsistenz). Verlangt mindestens intervallskalierte Urteile.	Sijtsma (2009); Wirtz/Caspar (2002)
ICC	Intraklassenkorrelation	Vgl. z. B. Shrout/Fleiss (1979)	$0 \leq \text{ICC} \leq 1$	Verallgemeinert den varianzanalytischen Ansatz nach Cronbachs Alpha. Es gibt mehrere Varianten der ICC, von denen einige nicht nur Konsistenz-, sondern auch Konsensinformation berücksichtigen. Sorgfältige Spezifizierung des varianzanalytischen Datenmodells erforderlich. Verlangt mindestens intervallskalierte Urteile.	McGraw/Wong (1996); Wirtz/Caspar (2002)

Tab. 1: Konsens- und Konsistenzmaße der Beurteilerreliabilität

		Konsensmaße				Konsistenzmaße			
Beurteiler	*N*	P_o	κ_w	RAI	OAI	*r*	τ_b	α	ICC(A,k)
07 / 10	20	.70	.67	.90	.78	.83	.78	.90	.89
13 / 16	20	.60	.67	.87	.92	.84	.84	.91	.91
12 / 03	20	.55	.29	.85	.61	.49	.42	.64	.65
17 / 11	19	.53	.42	.84	.67	.62	.58	.74	.75
14 / 08	23	.52	.50	.84	.63	.77	.70	.87	.82
08 / 12	24	.50	.54	.82	.80	.71	.64	.83	.83
09 / 17	26	.50	.34	.82	.65	.53	.49	.69	.64
05 / 18	21	.48	.53	.83	.72	.76	.68	.86	.85
02 / 04	24	.46	.33	.82	.53	.58	.52	.73	.70
10 / 09	21	.43	.41	.79	.61	.78	.72	.87	.76
15 / 07	28	.36	.20	.71	.59	.53	.48	.62	.48
13 / 03	21	.24	.22	.65	.50	.66	.62	.80	.57
05 / 07	20	.20	.22	.68	.40	.77	.72	.85	.63
01 / 14	20	.10	.00	.62	.22	.21	.26	.34	.15
M	–	.44	.38	.79	.62	.65	.60	.76	.69
SD	–	.16	.19	.09	.17	.17	.16	.15	.20

Tab. 2: Ergebnisse für Konsens- und Konsistenzmaße (exemplarische Datenanalyse)
Anmerkung: *N* = Anzahl der von je zwei unabhängigen Beurteilern bewerteten Teilnehmerleistungen im TestDaF-Subtest Schriftlicher Ausdruck. P_o = Exakte Übereinstimmung. κ_w = Gewichtetes Kappa. RAI = Rater Agreement Index. OAI = Overall Agreement Index. *r* = Produkt-Moment-Korrelation. τ_b = Kendalls Tau-b. α = Cronbachs Alpha. ICC(A,k) = Intraklassenkorrelation.

teils in anderen Untersuchungskontexten häufiger Verwendung finden. Auf eine formale Darstellung der einzelnen Maße wird aus Raumgründen verzichtet. Berechnungsvorschriften, vertiefende Diskussionen und Anwendungen finden sich in den zitierten Arbeiten. Je vier der in der Tabelle aufgeführten Konsens- bzw. Konsistenzmaße werden im nächsten Abschnitt anhand eines exemplarischen Datensatzes näher besprochen.

3 Eine exemplarische Reliabilitätsanalyse

Die hier verwendeten Daten stammten aus Beurteilungen von Leistungen im Schriftlichen Ausdruck als Teil der Sprachprüfung TestDaF (Test Deutsch als Fremdsprache; www.testdaf.de).[1] Jeder Schreibbogen wurde von zwei unabhängigen Beurteilern nach drei Kriterien (Gesamteindruck, Behandlung der Aufgabe, sprachliche Realisierung) anhand der TestDaF-Niveaustufen-Skala (TDN-Skala) mit vier Kategorien (unter TDN 3, TDN 3, TDN 4 und TDN 5) eingestuft. TDN 3 bis TDN 5 entsprechen dabei den Stufen B2.1 (selbstständige Sprachverwendung) bis C1.2 (kompetente Sprachverwendung) des Gemeinsamen europäischen Referenzrahmens für Sprachen (Europarat 2001; vgl. auch Kecker/Eckes 2010).

Für Zwecke der Datenanalyse wurde die Kategorie „unter TDN 3“ mit „2“ kodiert. Die Einzelbewertungen in den drei Kriterien wurden nach einem ungewichteten Mittelungsverfahren zu einer Gesamtbewertung zusammengefasst. Auf der Grundlage der Gesamtbewertungen wurden für die insgesamt 14 Paare von Beurteilern je vier Konsens- und Konsistenzkoeffizienten berechnet. Tab. 2 zeigt die Ergebnisse.

In der Tabelle sind die Beurteilerpaare nach der Höhe der exakten Übereinstimmung angeordnet. Die Übereinstimmung reicht von zufriedenstellenden 70 % für das Paar 07/10 bis zu inakzeptabel niedrigen 10 % für das Paar 01/14. Der gewichtete Kappa-Koeffizient ist für das Paar 01/14 gleich Null, was bedeutet, dass die beobachtete Übereinstimmung komplett durch Zufall erklärt werden kann.

[1] Die betrachtete TestDaF-Prüfung fand im Oktober 2001 in 29 Ländern statt. In die Analyse eingegangen sind Einstufungen der Leistung von 307 Teilnehmern durch 18 Beurteiler. Alle Beurteiler verfügten als Lehrer bzw. Prüfer über mehrjährige Berufserfahrung im Fach Deutsch als Fremdsprache und hatten eine gezielte Schulung im Hinblick auf diese Prüfung erhalten.

Den konventionellen Kappa-Grenzwert für eine „gute“ Übereinstimmung (Kappa = .60; vgl. z. B. Bortz/Döring 2006) übertreffen nur die Paare 07/10 und 13/16. Auch die beiden anderen Konsensmaße (RAI, OAI) nehmen für diese beiden Paare hohe Werte an. Beim RAI fällt auf, dass die Werte auf relativ hohem Niveau liegen und eine sehr geringe Streuung besitzen; d. h., dieser Index differenziert nur relativ schwach zwischen den Beurteilerpaaren.

Die Konsistenzkoeffizienten lassen ähnliche Schlüsse zu: Pearson-Korrelation, Kendalls Tau-b, Cronbachs Alpha und die ICC(A,k) sind bei den Paaren 07/10 und 13/16 zufriedenstellend, während die meisten anderen Paare weit davon entfernt sind, ausreichend hohe Reliabilität aufzuweisen.

Bei Cronbachs Alpha ist zu beachten, dass dieser Koeffizient formal identisch ist mit einer speziellen Variante der Intraklassenkorrelation, und zwar mit der ICC für die Schätzung der Konsistenz des „durchschnittlichen Beurteilers“. In der Terminologie von Wirtz/Caspar (2002) ist dies die justierte ICC für den Mittelwert über die Beurteiler ($ICC_{just,\ MW}$), in der Terminologie von McGraw/Wong (1996) handelt es sich um die ICC(C,k). Die in der rechten äußeren Spalte von Tab. 2 wiedergegebene ICC(A,k) berücksichtigt neben der Konsistenz auch den Konsens unter den Beurteilern (entspricht nach Wirtz/Caspar der unjustierten ICC für den Mittelwert über die Beurteiler; $ICC_{unjust,\ MW}$). Das hat zur Folge, dass die ICC(A,k) dann niedriger ausfällt als Cronbachs Alpha, wenn sich die Mittelwerte der Beurteiler unterscheiden (etwa im Falle von Unterschieden in der Beurteilerstrenge).

Ein Beispiel hierfür findet sich wieder in Tab. 2. Bei den Paaren 13/03 und 05/07 sind die Konsenswerte sehr niedrig, die Konsistenzwerte dagegen relativ hoch. Außerdem ist innerhalb der Gruppe der Konsistenzmaße bei diesen Paaren zu sehen, dass der Unterschied zwischen Cronbachs Alpha und ICC(A,k) so groß ist wie bei keinem der anderen Paare. Um das Zustandekommen dieser Differenzen aufzuklären, sind in Tab. 3 die Einstufungen, die Beurteiler 13 und 03 abgegeben haben, in Form einer Kreuztabelle einander gegenübergestellt.

Zehn der 16 Nichtübereinstimmungen betragen eine TDN-Stufe, die sechs übrigen betragen zwei TDN-Stufen. So hat z. B. Beurteiler 13 vier Teilnehmer nach TDN 3 eingestuft, Beurteiler 03 dieselben Teilnehmer aber nach TDN 5. Die Verteilung der Nichtübereinstimmungen scheint hierbei alles andere als zufällig. In keinem einzigen Fall hat Beurteiler 03

	Beurteiler 03				
Beurteiler 13	u. TDN 3	TDN 3	TDN 4	TDN 5	Zeilensumme
unter TDN 3	**1**	**3**	**2**		6
TDN 3			**5**	**4**	9
TDN 4			**2**	**2**	4
TDN 5				**2**	2
Spaltensumme	1	3	9	8	21

Tab. 3: Kreuzklassifikation der Einstufungen durch Beurteiler 13 und 03
Anmerkung: TDN = TestDaF-Niveaustufe.

	Beurteiler 16				
Beurteiler 13	u. TDN 3	TDN 3	TDN 4	TDN 5	Zeilensumme
unter TDN 3	**8**				8
TDN 3	**1**	**1**			2
TDN 4		**4**	**2**	**3**	9
TDN 5				**1**	1
Spaltensumme	9	5	2	4	20

Tab. 4: Kreuzklassifikation der Einstufungen durch Beurteiler 13 und 16
Anmerkung: TDN = TestDaF-Niveaustufe.

Beurteiler	Strenge	*SE*	Anzahl der Einzelurteile
16	2.40	0.30	60
13	2.09	0.20	123
15	1.21	0.22	84
04	0.29	0.23	72
17	–0.57	0.18	135
03	–2.01	0.19	123
07	–2.24	0.15	204

Tab. 5: Ausschnitt aus den Messergebnissen für die Beurteiler (exemplarische Datenanalyse)
Anmerkung: Die Messergebnisse stützen sich auf eine Multifacetten-Rasch-Analyse der Beurteilungen von Teilnehmerleistungen im TestDaF-Prüfungsteil Schriftlicher Ausdruck. Die Strenge der Beurteiler ist in Einheiten der Logitskala wiedergegeben. *SE* = Standardfehler.

die Teilnehmer niedriger eingestuft als Beurteiler 13. Die Tendenz von Beurteiler 03 geht damit eindeutig in die Richtung einer systematisch höheren (d. h. weniger strengen) Einstufung. Die Einstufungen von Beurteiler 03 sind gleichsam auf der TDN-Skala um eine bis zwei TDN-Stufen „nach oben" verschoben.

Ein ganz anderes Bild ergibt sich, wenn man die Einstufungen von Beurteiler 13 denjenigen von Beurteiler 16 gegenüberstellt (s. Tab. 4). Abweichende Einstufungen gibt es in nur acht Fällen. Die Abweichungen betragen jeweils nur eine TDN-Stufe. Das positive Bild bestätigen die Werte der Konsens- und Konsistenzmaße, die durchweg auf zufriedenstellendem Niveau liegen. Es drängt sich der Schluss auf, Beurteiler 13 und 16 hätten die Leistungen der 20 Teilnehmer hinreichend reliabel eingestuft. An der Berechtigung dieses Schlusses sollte sich kaum ein Zweifel regen – es sei denn, ein genauerer Blick auf die Urteilstendenzen der Beurteiler führte zu einer begründet anderen Einschätzung. Wie es zu einer solchen Einschätzung kommen könnte, zeigt der folgende Abschnitt.

4 *Strenge, Milde und das Übereinstimmungs-Genauigkeits-Paradox*

Die folgende Analyse des exemplarischen Datensatzes stützte sich auf ein psychologisches Testmodell, das so genannte Multifacetten-Rasch-Modell (vgl. Linacre 1989).[1] Dieses Modell erlaubt es, neben Teilnehmern und Items (Aufgaben) auch Beurteiler, Kriterien, Urteilszeitpunkte und weitere interessierende Variable in ihrem Einfluss auf Beurteilungen zu untersuchen. Die potenziell Einfluss nehmenden Variablen heißen „Facetten" der Test- oder Urteilssituation (vgl. Bachman 2004; Eckes 2005; 2009; 2011).

Im vorliegenden Fall lieferte das Modell bzw. das zu seiner Implementierung verwendete Computerprogramm FACETS (vgl. Linacre 2010) Messungen der Fähigkeit der Teilnehmer, der Strenge bzw. Milde der Beurteiler und der Schwierigkeit der Kriterien auf einer gemeinsamen linearen Skala (der Logitskala). Tab. 5 zeigt einen Ausschnitt aus den Messergebnissen für die Beurteiler (eine Darstellung der Messergebnisse für alle Beurteiler findet sich in Eckes 2009; 2011). Die Strenge der Beurteiler ist in Einheiten der Logitskala angegeben. Positive Logitwerte verweisen auf Strenge, negative Logitwerte auf Milde. Der Standardfehler (*SE*) gibt die Präzision der Schätzung des Strengeparameters wieder.

Die Strengedifferenzen sind beträchtlich. Der strengste Beurteiler ist Beurteiler 16 mit 2.40 Logits, der mildeste ist Beurteiler 07 mit –2.24 Logits. In Einheiten der TDN-Skala ausgedrückt beträgt die Differenz zwischen diesen beiden Beurteilern mehr als eine ganze TDN-Stufe (TDN-Differenz = 1.23). Es sei angemerkt, dass die hier festgestellten Unterschiede in der Beurteilerstrenge für Verfah-

[1] Das Multifacetten-Rasch-Modell gehört zur Familie der Rasch-Modelle (vgl. Bond/Fox 2007; Rost 2004). Diese Modelle sind nach dem dänischen Mathematiker und Statistiker Georg Rasch (1901–1980) benannt. Rasch-Modelle spielen in der Analyse und Evaluation von pädagogisch-psychologischen Testverfahren eine herausragende Rolle. Z. B. bilden sie die psychometrische Grundlage für internationale Schulleistungsvergleiche wie PISA, PIRLS oder TIMSS (vgl. z. B. Hartig 2007; Seidel/Prenzel 2008).

ren der beurteilergestützten Leistungsmessung im Bereich von Sprachleistungen wie auch in vielen anderen Leistungsbereichen ganz und gar typisch sind (vgl. Eckes 2010).

Vor dem Hintergrund der in Tab. 5 gezeigten Ergebnisse der Multifacetten-Rasch-Analyse stellt sich die Frage nach der „Genauigkeit" der Einstufungen durch Beurteiler 13 und 16 in neuem Licht. Diese Beurteiler waren die beiden strengsten in der Beurteilergruppe überhaupt. Es ist daher alles andere als eine Überraschung, dass Beurteiler 13 und 16 ähnlich einstuften, unterlagen sie doch ähnlich stark ausgeprägten Strengetendenzen. Mit anderen Worten: Beide Beurteiler unterschätzten die Fähigkeit der Teilnehmer (bezogen auf die Gruppe der untersuchten Beurteiler).

Beurteiler 03 gehörte mit –2.01 Logits zu den mildesten Beurteilern. Dies bestätigt die oben diskutierte Urteilstendenz (s. Tab. 3). Zugleich wird deutlich, dass angesichts der großen Strengedifferenz zwischen Beurteiler 03 und Beurteiler 13 keine hohe Übereinstimmung in den jeweiligen Einstufungen zu erwarten war. Während Beurteiler 13 also die Fähigkeit der Teilnehmer unterschätzte, unterlag Beurteiler 03 der gegenläufigen Tendenz, die Fähigkeit der Teilnehmer zu überschätzen.

Die exemplarische Multifacetten-Rasch-Analyse verweist auf ein fundamentales Problem des traditionellen Ansatzes zur Ermittlung der Beurteilerreliabilität als Methode des Nachweises von Urteilsgenauigkeit: Hohe Übereinstimmung zwischen Beurteilern impliziert nicht notwendig hohe Genauigkeit der Beurteilungen. Empirisch nachgewiesene hohe Beurteilerreliabilität kann zu dem Fehlschluss verleiten, Beurteiler lieferten genaue Einstufungen, wenn tatsächlich eine substanzielle Unterschätzung oder eine substanzielle Überschätzung der Fähigkeit von Teilnehmern vorliegt. Dies ist im Kern das Übereinstimmungs-Genauigkeits-Paradox (vgl. Eckes 2009; 2010).

Wird Beurteilerreliabilität mittels einer Konsensmethode bestimmt, so sind hohe Übereinstimmungsraten allein schon dann zu erwarten, wenn die betrachteten Beurteiler ähnlich streng oder ähnlich milde urteilen. Die Frage der Urteilsgenauigkeit lässt sich so nicht beantworten. Auch die Berechnung von Konsistenzkoeffizienten hilft nicht weiter, da ähnlich streng (oder ähnlich milde) einstufende Beurteiler weitgehend übereinstimmende Rangreihen von Teilnehmerleistungen erstellen. Umgekehrt darf aus niedrigen Konsens- und Konsistenzwerten nicht der Schluss gezogen werden, die beteiligten Beurteiler lieferten ungenaue Urteile. Betrachten wir hierzu noch einmal das Beurteilerpaar 01/14 (s. Tab. 2). Die Werte für dieses Paar liegen durchweg so niedrig, dass man daran denken könnte, Beurteiler 01 und Beurteiler 14 von weiteren Beurteilungen auszuschließen. Tatsächlich aber zeigt die Multifacetten-Rasch-Analyse, dass die niedrige Beurteilerreliabilität primär auf große Strengedifferenzen zurückgeht (Beurteiler 01: –2.23 Logits, Beurteiler 14: 1.83 Logits; vgl. ausführlicher Eckes 2009).

In ihren vielzitierten Arbeiten über Maße der Beurteilerreliabilität vertraten Tinsley/Weiss (1975; 2000) die Auffassung, dass im Falle niedriger Konsenswerte und niedriger Konsistenzwerte die Beurteilungen invalide und für Forschungszwecke oder für angewandte Fragestellungen ungeeignet seien. Diese Auffassung ist in ihrer Allgemeinheit klar zurückzuweisen. Erst eine Multifacetten-Rasch-Analyse kann letztlich Aufschluss darüber geben, wie verlässlich und aussagekräftig die von Beurteilern abgegebenen Einstufungen sind.

5 Empfehlungen

Die Beurteilung fremdsprachlicher Leistung ist ein komplexer Prozess, der vielen Fehlereinflüssen unterliegt. Hierzu zählen insbesondere Fehler auf Seiten der Beurteiler. Es gilt, diese Fehler so weit wie möglich zu verringern, zu kontrollieren und in ihren Wirkungen zu erfassen. Die Bestimmung der Beurteilerreliabilität ist hierbei von grundlegender Bedeutung. Zugleich aber bedürfen Auswahl und Anwendung von Methoden der Reliabilitätsbestimmung sorgfältiger Überlegung, um bei der Evaluation der psychometrischen Qualität von Beurteilungen Fehlschlüsse zu vermeiden.

Das Datenbeispiel hat gezeigt, dass eine Multifacetten-Rasch-Analyse wertvolle Einsichten in die Qualität von Leistungsbeurteilungen vermitteln kann, und zwar auch bei relativ geringer Anzahl von Beurteilern, Kriterien und Teilnehmern. In der Literatur sind derartige Analysen auch noch bei deutlich kleineren Datensätzen dokumentiert (vgl. Eckes 2010). Als grobe Orientierung kann dienen, dass pro Kategorie einer Ratingskala mindestens 10 Datenpunkte oder Be-

obachtungen (z. B. Urteile) vorliegen sollten (vgl. Linacre 2004). Weiter ist vorteilhaft, dass Multifacetten-Rasch-Analysen gegenüber fehlenden Werten relativ unempfindlich sind. Einen geeigneten Beurteilungsplan vorausgesetzt, reicht es schon aus, wenn jeder Teilnehmer nur von einem einzigen Beurteiler eingestuft wird (vgl. Eckes 2009). Gegenüber einem faktoriellen Plan, in dem alle Teilnehmer von allen Beurteilern eingestuft werden, kann dies eine Ersparnis von über 90 % der Beobachtungen mit sich bringen.

Wenn eine Multifacetten-Rasch-Analyse nicht möglich ist, aber eine Aussage über die Beurteilerreliabilität getroffen werden soll, sind wenigstens ein Konsensmaß und ein Konsistenzmaß zu bestimmen. Welche Maße jeweils in Frage kommen, hängt wesentlich von den Eigenschaften der Daten und dem Verwendungszweck ab. Tab. 1 enthält hierzu einige Hinweise.

Zeichnen sich deutliche Unterschiede zwischen Konsens- und Konsistenzwerten ab, so deutet dies darauf hin, dass innerhalb der betrachteten Gruppe von Beurteilern nicht vernachlässigbare Strengedifferenzen vorliegen. Ähnliche Erkenntnisse kann auch ein Vergleich zwischen Cronbachs Alpha und der ICC(A,k) liefern. Um nicht bei alleiniger Betrachtung von Konsens- und/oder Konsistenzmaßen in die Irre zu gehen, ist stets ein ergänzender Blick auf die Rohdaten etwa in Form von Kreuztabellen oder Streuungsdiagrammen ratsam.

Literatur

Bachman, Lyle F. (2004): Statistical analyses for language assessment. Cambridge.

Bond, Trevor G. / Fox, Christine M. (2007): Applying the Rasch model. Fundamental measurement in the human sciences. 2. Aufl. Mahwah, NJ.

Bortz, Jürgen / Döring, Nicola (2006): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 4. Aufl. Berlin.

Bortz, Jürgen u. a. (2000): Verteilungsfreie Methoden in der Biostatistik. 2. Aufl. Berlin.

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

Burke, Michael J. u. a. (1999): On average deviation indices for estimating interrater agreement. In: Organizational Research Methods 2, 49–68.

Burry-Stock, Judith A. u. a. (1996): Rater agreement indexes for performance assessment. In: Educational and Psychological Measurement 56, 251–262.

Cohen, Ayala u. a. (2009): Testing agreement for multi-item scales with the indices $r_{WG(J)}$ and $AD_{M(J)}$. In: Organizational Research Methods 12, 148–164.

Cohen, Jacob (1960): A coefficient of agreement for nominal scales. In: Educational and Psychological Measurement 20, 37–46.

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

Cronbach, Lee J. (1951): Coefficient alpha and the internal structure of tests. In: Psychometrika 16, 297–334.

Dunlap, William P. (2003): Accurate tests of statistical significance for r_{WG} and average deviation interrater agreement indexes. In: Journal of Applied Psychology 88, 356–362.

Eckes, Thomas (2005): Analyse und Evaluation sprachproduktiver Prüfungen beim TestDaF. In: I. Kühn u. a. (Hg.), Sprachtests in der Diskussion. Frankfurt a. M. u. a., 60–93 (Wittenberger Beiträge zur deutschen Sprache und Kultur, 4).

Eckes, Thomas (2009): Many-facet Rasch measurement. In: S. Takala (Hg.), Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H). Strasbourg: Council of Europe / Language Policy Division. Online: http://www.coe. int/t/dg4/linguistic/manuel1_EN.asp?#P19_2121.

Eckes, Thomas (2010): Die Beurteilung sprachlicher Kompetenz auf dem Prüfstand. Fairness in der beurteilergestützten Leistungsmessung. In: K. Schramm u. a. (Hg.), Fremdsprachliches Handeln beobachten, messen und evaluieren. Neue methodische Ansätze der Kompetenzforschung und Videographie. Frankfurt a. M. u. a., 65–97.

Eckes, Thomas (2011): Introduction to many-facet Rasch measurement. Analyzing and evaluating rater-mediated assessments. Frankfurt a. M. u. a.

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

Europarat (2001): Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen. Berlin u. a.

von Eye, Alexander / Mun, Eun Y. (2005): Analyz-

ing rater agreement. Manifest variable methods. Mahwah, NJ.
Fleiss, Joseph L. u. a. (2003): Statistical methods for rates and proportions. 3. Aufl. Hoboken, NJ.
Grotjahn, Rüdiger (2000): Testtheorie. Grundzüge und Anwendungen in der Praxis. In: A. Wolff / H. Tanzer (Hg.), Sprache – Kultur – Politik. Regensburg, 304–341.
Grotjahn, Rüdiger (2007): Testen und Prüfen. Aktuelle Tendenzen. In: Neue Beiträge zur Germanistik 6, 19–38.
Hamp-Lyons, Liz (2007): Worrying about rating. In: Assessing Writing 12, 1–9.
Hartig, Johannes (2007): Skalierung und Definition von Kompetenzniveaus. In: B. Beck / E. Klieme (Hg.), Sprachliche Kompetenzen. Konzepte und Messung. Weinheim, 83–99.
Hayes, Andrew F. / Krippendorff, Klaus (2007): Answering the call for a standard reliability measure for coding data. In: Communication Methods and Measures 1, 77–89.
James, Lawrence R. u. a. (1984): Estimating within-group interrater reliability with and without response bias. In: Journal of Applied Psychology 69, 85–98.
Jones, Neil (2005): Seminar to calibrate examples of spoken performance. Report on analysis of rating data. Strasbourg: Council of Europe / Language Policy Division. Online: http://www.coe.int/T/DG4/Portfolio/documents/Sevresreport NJ.pdf.
Kecker, Gabriele / Eckes, Thomas (2010): Putting the Manual to the test. The TestDaF–CEFR linking project. In: W. Martyniuk (Hg.), Aligning tests with the CEFR. Reflections on using the Council of Europe's draft Manual. Cambridge, 50–79.
Kendall, Maurice G. (1948): Rank correlation methods. London.
LeBreton, James M. / Senter, Jenell L. (2008): Answers to 20 questions about interrater reliability and interrater agreement. In: Organizational Research Methods 11, 815–852.
Linacre, John M. (1989): Many-facet Rasch measurement. Chicago.
Linacre, John M. (2004): Optimizing rating scale category effectiveness. In: E. V. Smith / R. M. Smith (Hg.), Introduction to Rasch measurement. Maple Grove, MN, 258–278.
Linacre, John M. (2010): Facets Rasch measurement computer program [Computer software]. Chicago.
Lüdtke, Oliver / Robitzsch, Alexander (2009): Assessing within-group agreement. A critical examination of a random-group resampling approach. In: Organizational Research Methods 12, 461–487.
McGraw, Kenneth O. / Wong, Seok P. (1996): Forming inferences about some intraclass correlation coefficients. In: Psychological Methods 1, 30–46.
McNamara, Tim F. (1996): Measuring second language performance. London.
Moosbrugger, Helfried / Höfling, Volkmar (2007): Standards für psychologisches Testen. In: H. Moosbrugger / A. Kelava (Hg.), Testtheorie und Fragebogenkonstruktion. Heidelberg, 193–212.
Mun, Eun Y. (2005): Rater agreement – weighted kappa. In: B. S. Everitt / D. Howell (Hg.), Encyclopedia of statistics in behavioral science. Bd. 3. New York, 1714–1715.
Murphy, Kevin R. / DeShon, Richard (2000): Interrater correlations do not estimate the reliability of job performance ratings. In: Personnel Psychology 53, 873–900.
Rost, Jürgen (2004): Lehrbuch Testtheorie, Testkonstruktion. 2. Aufl. Bern.
Seidel, Tina / Prenzel, Manfred (2008): Assessment in large-scale studies. In: J. Hartig u. a. (Hg.), Assessment of competencies in educational contexts. Göttingen, 279–304.
Shoukri, Mohamed M. (2004): Measures of interobserver agreement. Boca Raton, FL.
Shrout, Patrick E. / Fleiss, Joseph L. (1979): Intraclass correlations. Uses in assessing rater reliability. In: Psychological Bulletin 86, 420–428.
Sijtsma, Klaas (2009): On the use, the misuse, and the very limited usefulness of Cronbach's alpha. In: Psychometrika 74, 107–120.
Stemler, Steven E. / Tsai, Jessica (2008): Best practices in interrater reliability. Three common approaches. In: J. W. Osborne (Hg.), Best practices in quantitative methods. Los Angeles, 29–49.
Tinsley, Howard E. A. / Weiss, David J. (1975): Interrater reliability and agreement of subjective judgments. In: Journal of Counseling Psychology 22, 358–376.
Tinsley, Howard E. A. / Weiss, David J. (2000): Interrater reliability and agreement. In: H. E. A. Tinsley / S. D. Brown (Hg.), Handbook of applied multivariate statistics and mathematical modeling. San Diego, CA, 95–124.
Uebersax, John (2008): Statistical methods for rater agreement. Online: http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm.
Wigglesworth, Gillian (2008): Task and performance based assessment. In: E. Shohamy / N. H. Hornberger (Hg.), Encyclopedia of language and education. Bd. 7: Language testing and assessment. 2. Aufl. New York, 111–122.
Wirtz, Markus (2006): Methoden zur Bestimmung der Beurteilerübereinstimmung. In: F. Petermann / M. Eid (Hg.), Handbuch der Psychologischen Diagnostik. Göttingen, 369–380.
Wirtz, Markus / Caspar, Franz (2002): Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen. Göttingen.
Zegers, Frits E. (1991): Coefficients for interrater agreement. In: Applied Psychological Measurement 15, 321–333.