Qualitätssicherung beim TestDaF: Konzepte, Methoden, Ergebnisse

Thomas Eckes (Hagen)

1. Notwendigkeit einer testmethodischen Qualitätssicherung

Seit seiner Einführung im Jahr 2001 haben am „Test Deutsch als Fremdsprache“ (TestDaF) über 11.000 Personen an mehr als 240 Testzentren in 72 Ländern teilgenommen (Stand: November 2003). Im Jahr 2004 ist mit einem nochmals deutlichen Anstieg der Teilnehmerzahlen zu rechnen. Einen wesentlichen Beitrag zum Erfolg des TestDaF im In- und Ausland leistet seine hohe testmethodische Qualität. Wie diese Qualität erreicht und dauerhaft gesichert wird, ist Thema der vorliegenden Arbeit.

Getrennt nach den vier Fertigkeiten Leseverstehen, Hörverstehen, Schriftlicher Ausdruck und Mündlicher Ausdruck informieren die zentral vom TestDaF-Institut ausgestellten Zeugnisse darüber, welche Kompetenzstufen die Prüfungsteilnehmer[1] erreicht haben. Jede der Stufen (TestDaF-Niveaustufen, kurz TDN-Stufen) ist auf der Rückseite des Zeugnisses in Form von Kann-Beschreibungen näher erläutert (siehe auch unter www.testdaf.de). Diese Beschreibungen orientieren sich an den „Can-Do-Skalen“ der *Association of Language Testers in Europe* (ALTE) bzw. den Kann-Beschreibungen des *Gemeinsamen europäischen Referenzrahmens* (vgl. Europarat, 2001). So erhalten die Teilnehmer genauen Aufschluss darüber, in welchen Fertigkeitsbereichen ihre Stärken bzw. Schwächen liegen. Zugleich können Hochschulen, denen ein TestDaF-Zeugnis vorgelegt wird, anhand der differenzierten Feststellung des Sprachniveaus eine fundierte Entscheidung über die Zulassung oder Ablehnung eines Bewerbers treffen.

Da es sich beim TestDaF um einen Sprachtest handelt, mit dessen Ergebnissen für die Teilnehmer weitreichende persönliche Konsequenzen im Sinne eines „High-Stakes-Tests“ verbunden sind, kommt es darauf an sicherzustellen, dass Konstruktion, Analyse und Evaluation des TestDaF hohen wissenschaftlichen Qualitätskriterien entsprechen. Derartige Kriterien sind z.B. in den „Standards für pädagogisches und psychologisches Testen“ festgehalten (Häcker, Leutner & Amelang, 1998; vgl. auch American Educational Research Association, 1999).

---

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

Fremdsprachen und Hochschule, 2003, Heft 69, S. 43–68.

Im Folgenden werden zentrale Aspekte der testmethodischen Qualitätssicherung beim TestDaF vorgestellt und anhand exemplarischer Evaluationsergebnisse erläutert. Zunächst wird skizziert, welchen Regeln die Konstruktion einer Aufgabensammlung, die bei einer späteren TestDaF-Prüfung zum Einsatz kommen soll, unterliegt. Anschließend werden verschiedene Methoden der Qualitätssicherung in den rezeptiven und produktiven Teilprüfungen des TestDaF genauer besprochen. In einem weiteren Abschnitt kommen besondere Probleme der Qualitätssicherung von Leistungsbewertungen im Schriftlichen und Mündlichen Ausdruck zur Sprache.

2. Konstruktion und Evaluation einer TestDaF-Aufgabensammlung

2.1 Der TestDaF-Erprobungszyklus

Bevor eine neu erstellte Aufgabensammlung im Rahmen einer TestDaF-Prüfung verwendet werden kann, müssen alle darin enthaltenen Subtests, Aufgaben und Items nach verschiedenen testmethodischen Kriterien analysiert und im Detail evaluiert werden. Oberstes Ziel dabei ist, eine möglichst hohe Objektivität, Reliabilität und Validität unter Einschluss der Fairness von Leistungsbewertungen zu gewährleisten.[2] Um eine gegebene Aufgabensammlung im Hinblick auf diese Ziele zu optimieren, werden ihre verschiedenen Komponenten einer sorgfältigen Qualitätskontrolle unterworfen. Diese Kontrolle vollzieht sich in mehreren aufeinander folgenden Evaluationsschritten, die in die Phasen *Vorerprobung*, *Revision* und *Erprobung* unterteilt sind. Der gesamte Evaluationsprozess, der *TestDaF-Erprobungszyklus*, ist in Abbildung 1 schematisch dargestellt.[3]

---

[2] Eine ausgezeichnete Diskussion der psychometrischen Gütekriterien im Kontext des Sprachtestens findet sich in Grotjahn (2000). Vertiefende Darstellungen der testtheoretischen Grundlagen geben z.B. Bortz und Döring (2002, Kap. 4), Moosbrugger (1999) und Rost (1996).

[3] Die für die Erstellung eines vorläufigen Satzes von Aufgaben und Items notwendigen Überlegungen und Schritte sind dabei nicht berücksichtigt (für nähere Informationen hierzu siehe Arras & Grotjahn, 2002; Kniffka & Üstünsöz-Beurer, 2001; Projektgruppe TestDaF, 2000).

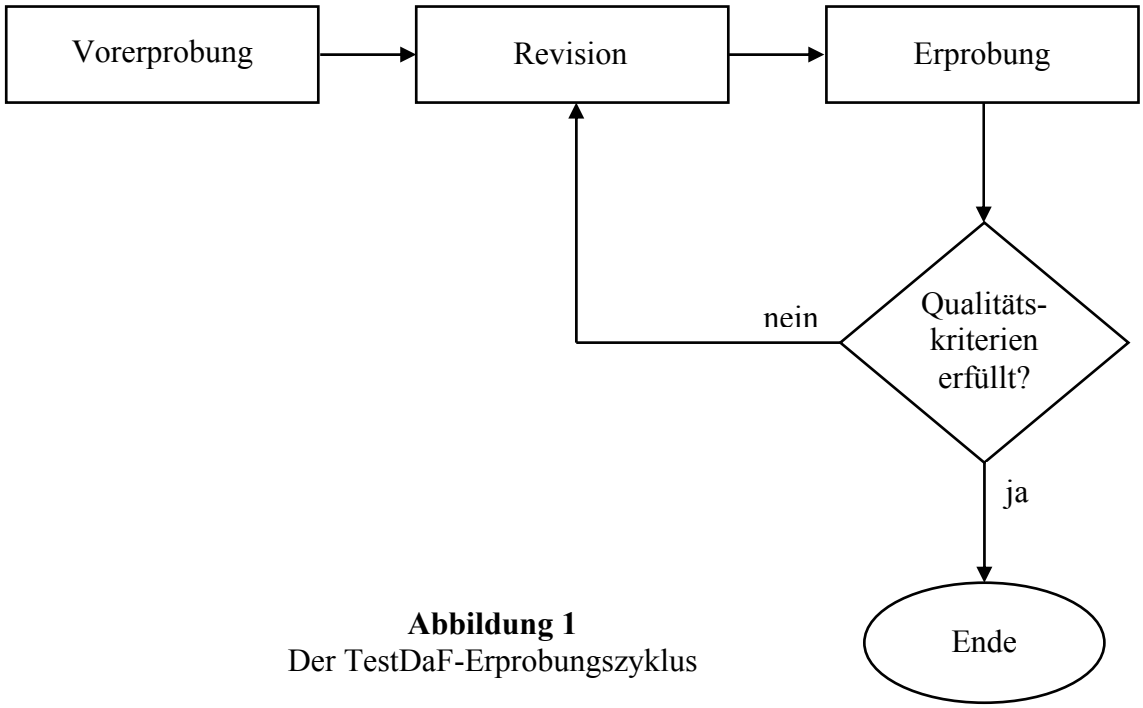

Abbildung 1
Der TestDaF-Erprobungszyklus

2.2 Vorerprobung

Die Vorerprobung hat den Zweck, eine neu entwickelte Aufgabensammlung einer ersten, noch recht groben testmethodischen Analyse auf der Grundlage der klassischen Testtheorie (vgl. z.B. Lienert & Raatz, 1998) zu unterziehen. Als Datenbasis dienen zum einen die Antworten bzw. Lösungen der Teilnehmer (im Folgenden auch Probanden oder Pbn) zu den Aufgaben in den beiden rezeptiven Subtests Leseverstehen (3 Aufgaben mit je 10 Items) und Hörverstehen (3 Aufgaben mit 8, 10 bzw. 7 Items), zum anderen die Bewertungen der Leistungen in den beiden produktiven Subtests Schriftlicher Ausdruck (1 Aufgabe) und Mündlicher Ausdruck (9 Aufgaben) durch TestDaF-geschulte Beurteiler[4]. Darüber hinaus erhalten die Pbn einen Fragebogen, anhand dessen sie einzelne Aufgaben bzw. Items z.B. nach wahrgenommener Schwierigkeit oder nach Klarheit der Aufgabenstellung einschätzen können.

In der Vorerprobung werden sowohl Fremdsprachler (hauptsächlich an inländischen Testzentren, ca. 40 – 60 Pbn) als auch Muttersprachler (an inländischen Testzentren, ca. 20 Pbn) untersucht. Das Hauptaugenmerk liegt dabei auf den Lösungsraten, die für die dichotomen (d.h. entweder als

[4] Der Subtest Mündlicher Ausdruck enthält noch eine Aufwärmaufgabe, die aber nicht in die Auswertung bzw. Niveaustufen-Feststellung einbezogen ist.

„richtig“ oder als „falsch“ kodierten) Items in den Subtests Leseverstehen und Hörverstehen ermittelt werden. Die *Lösungsrate* eines gegebenen Items ist dabei definiert als der Anteil der Pbn, die das Item richtig beantworten. Liegen die Lösungsraten bei den Fremdsprachlern zu niedrig oder zu hoch (d.h. sind einzelne Items für diese Pbn-Gruppe zu schwierig oder zu leicht), dann werden die betreffenden Items für eine Revision vorgesehen. Das Gleiche gilt für Items, die wenigstens 10% der Muttersprachler nicht korrekt beantwortet haben.

Darüber hinaus finden im Falle der Fremdsprachler die Itemtrennschärfen und die Lösungsraten der Distraktoren Eingang in die Qualitätsbeurteilung. Die *Trennschärfe* eines Items ist ein Maß dafür, wie gut die Testwerte der Pbn (ermittelt als die Anzahl der richtig beantworteten Items) aufgrund der Antworten der Pbn auf das gegebene Item vorhergesagt werden können. Bei den Mehrfachwahl-Items gibt eine *Distraktorenanalyse* Aufschluss darüber, inwieweit die falschen Antwortalternativen leistungsschwache Pbn ansprechen. Die Antworten der fremdsprachlichen Teilnehmer zu den Aufgaben im Schriftlichen und Mündlichen Ausdruck werden inhaltlich genau untersucht, um nichtintendierte Einflüsse auf die Aufgabenschwierigkeiten, wie z.B. missverständliche Fragen oder unklare Formulierungen, zu identifizieren. Ergänzt werden die Aufgabenanalysen in den produktiven Testteilen durch eine Analyse der beobachteten Verteilungen von TestDaF-Niveaustufen.

2.3 Revision

Im Anschluss an die Vorerprobung werden alle Aufgaben und Items, die im Verlaufe der testmethodischen Auswertungen als in irgendeiner Weise problematisch eingestuft wurden, einer sorgfältigen Revision unterzogen. Je nach Art und Umfang der diagnostizierten Schwachstellen werden ganze Aufgaben oder auch nur einzelne Items modifiziert oder durch neue ersetzt. Die Revision umfasst prinzipiell alle Elemente der zu prüfenden Aufgabensammlung, die in kritischer Weise Einfluss auf die Qualität von Aufgaben oder Items nehmen, also z.B. auch die Reihenfolge der Distraktoren bei Items des Leseverstehens, die Sprechgeschwindigkeit des Sprechers beim Hörverstehen oder die grafische Gestaltung von Diagrammen im Schriftlichen Ausdruck.

2.4 Erprobung

Die vorerprobte und gründlich überarbeitete Aufgabensammlung wird in der Erprobungsphase einer erneuten, weit umfangreicheren und differenzierteren testmethodischen Analyse unterzogen. Erprobungen sind das Herzstück der testmethodischen Qualitätskontrolle beim TestDaF. Anders als in der Vorerprobung werden hier ausschließlich Fremdsprachler (ca. 200-250 Pbn) an überwiegend ausländischen Testzentren untersucht. Zu diesem Zweck werden TestDaF-Testzentren im Lizenzierungsverfahren vertraglich darauf verpflichtet, an Erprobungsprüfungen teilzunehmen. Für sämtliche Pbn soll gelten, dass sie zur Zielgruppe der Teilnehmer an einer offiziellen TestDaF-Prüfung gehören, d.h. vor allem, dass sie hinreichende Deutschkenntnisse besitzen (empfohlen werden mindestens 700 Stunden Deutschunterricht).

Die Hauptziele von TestDaF-Erprobungsprüfungen sind: (a) eine detaillierte Untersuchung der psychometrischen Qualität der beiden rezeptiven Teilprüfungen Leseverstehen und Hörverstehen, (b) eine eingehende Analyse der Leistungsbeurteilungen in den beiden produktiven Teilprüfungen Schriftlicher Ausdruck und Mündlicher Ausdruck und (c) eine Erfassung zusätzlicher Sprachtestdaten (durch einen C-Test) zum Zwecke der Verankerung der Itemschwierigkeiten in den rezeptiven Teilprüfungen.

In der Regel bearbeiten die Teilnehmer an einer Erprobungsprüfung insgesamt fünf Prüfungsteile, die wie folgt angeordnet sind: Leseverstehen (Dauer 60 min), C-Test (20 min), Hörverstehen (40 min), Schriftlicher Ausdruck (60 min), Mündlicher Ausdruck (30 min). In Ausnahmefällen können sich Testzentren, die sich aufgrund räumlicher oder zeitlicher Engpässe nicht in der Lage sehen, alle fünf Einzelprüfungen durchzuführen, auf die Administration der rezeptiven Teilprüfungen (einschl. C-Test) oder auf die Administration der produktiven Teilprüfungen beschränken.

Den Schwerpunkt der Analyse von Erprobungsdaten bildet die Anwendung von Testmodellen, die im Detail festzustellen erlauben, wie gut eine gegebene Aufgabensammlung die jeweiligen Sprachfertigkeiten erfasst. Im Falle von signifikanten Modellabweichungen wird die betreffende Aufgabensammlung einer gezielten Revision unterzogen. Die revidierte Aufgabensammlung wird danach in einer erneuten Erprobungsprüfung im Hinblick auf die Erreichung der Qualitätsstandards evaluiert. Erfüllt eine Aufgabensammlung die testmethodischen Kriterien, endet der TestDaF-Erprobungszyklus. Die so empirisch bewährte Aufgabensammlung kann für

einen Einsatz in einer offiziellen TestDaF-Prüfung bereitgestellt werden. Um darüber hinaus das gleiche Schwierigkeitsniveau der Items im Leseverstehen bzw. Hörverstehen in verschiedenen TestDaF-Prüfungen zu gewährleisten, erfolgt eine Verankerung der Itemschwierigkeiten auf der Grundlage einer Skalierung von C-Test-Items. Wie bei der Evaluation von Erprobungsprüfungen und bei der Verankerung im Einzelnen vorgegangen wird, ist Gegenstand des nachfolgenden Abschnitts.

3. Qualitätssicherung in TestDaF-Erprobungsprüfungen

3.1 Item-Response-Theorie

Den theoretischen Hintergrund für eine Qualitätssicherung im Rahmen von Erprobungsprüfungen bilden Testmodelle, die unter der Bezeichnung *Item-Response-Theorie* (IRT) zusammengefasst werden (vgl. z.B. Embretson & Reise, 2000; Fischer, 1974, 1996; Hambleton, Robin & Xing, 2000; Rost, 1996; Steyer & Eid, 2001). IRT-Modelle enthalten formalisierte Annahmen über das Antwortverhalten, das Personen bei der Bearbeitung einzelner Testaufgaben bzw. Items zeigen. Die Grundannahme dabei lautet, dass die Antworten auf ein Testitem durch nicht direkt beobachtbare, latente Merkmale (Dimensionen, Variablen) erklärt werden können. In den meisten Anwendungsfällen wird postuliert, dass ein einziges Merkmal der Testperson (z.B. ihre Sprachfähigkeit) darauf Einfluss nimmt, mit welcher Wahrscheinlichkeit sie das Item löst (bei dichotomen Items) bzw. eine ganz bestimmte von mehreren geordneten Antwortkategorien wählt (bei polytomen Items). Diese und einige weitere Annahmen werden in Form eines mathematischen Modells ausgedrückt, das präzise Vorhersagen darüber erlaubt, wie Testpersonen mit unterschiedlichen Ausprägungen auf der latenten Dimension auf ein bestimmtes Item antworten. IRT-Modelle haben sich in den letzten 10 bis 15 Jahren zur dominierenden Methodologie im Bereich der Konstruktion, Analyse und Evaluation standardisierter Sprachprüfungen entwickelt (vgl. z.B. Bachman, 2000; Hambleton et al., 2000).

Bei der IRT-Analyse von TestDaF-Erprobungsdaten kommen Vertreter einer bestimmten Klasse von IRT-Modellen, die sog. *Rasch-Modelle* (Rasch, 1960/1980), zur Anwendung. Rasch-Modelle besitzen eine Reihe von psychometrischen Vorzügen, die sich im vorliegenden Kontext in fünf Punkten zusammenfassen lassen.

(1) Gilt das Rasch-Modell in der untersuchten Population von Personen bzw. Items, dann kann die Differenz zweier Personenparameter (z.B. die Differenz aus den Messungen der Sprachfähigkeit zweier Personen) unabhängig davon bestimmt werden, welche Items für den Vergleich herangezogen werden und welche Fähigkeiten die anderen getesteten Personen haben. Umgekehrt kann bei Modellgültigkeit die Differenz zweier Itemparameter unabhängig davon bestimmt werden, welche Personen untersucht werden und welche Schwierigkeiten die anderen Testitems aufweisen.

(2) Gilt das Rasch-Modell, dann schöpft der Testwert, in der Regel ermittelt als die Anzahl der gelösten oder richtig beantworteten Items, die gesamte Information aus, die ein Antwortmuster über die Ausprägung der latenten Personenvariablen enthält. Kubinger (1999) hat betont, dass Testwerte (wie oben definiert) nur dann die empirischen Verhaltens- bzw. Leistungsrelationen adäquat abbilden, d.h. *verrechnungsfair* sind, wenn das Rasch-Modell gilt. Verrechnungsfairness setzt also die Gültigkeit des Rasch-Modells notwendig voraus. Testwerte sind z.B. dann nicht verrechnungsfair, wenn leichte Items leistungsschwächere Personen gegenüber leistungsstärkeren Personen bevorteilen.

(3) Rasch-Modelle liefern für jede einzelne Parameterschätzung einen eigenen (standardisierten) Fehlerwert als Maß für die *Genauigkeit* der Schätzung. Dieser Standardfehler sinkt bei steigender Anzahl von Beobachtungen, die in die Parameterschätzung eingehen, und steigt bei Parameterschätzungen, die sich am unteren oder oberen Spektrum des Fähigkeitskontinuums bewegen. Anhand der Information über die Größe des jeweiligen Standardfehlers ist es z.B. möglich, Unterschiede zwischen zwei beliebigen Parameterwerten auf statistische Signifikanz zu testen.

(4) Die Analyse von Tests auf der Basis von Rasch-Modellen erlaubt einen quantitativen Vergleich zwischen beobachteten und erwarteten Antworten. Dabei sind die erwarteten Antworten die aufgrund des jeweiligen Modells vorhergesagten Antworten. Geringe Abweichungen zwischen beobachteten und erwarteten Antworten, d.h. kleine (standardisierte) *Residuen*, geben einen Hinweis darauf, dass die Daten gut mit den Annahmen des Modells übereinstimmen. Große Werte der Residuen lassen auf eine unzureichende Passung zwischen Modell und Daten schließen.

(5) Rasch-Modelle unterstützen in effizienter Weise die Entwicklung und Anwendung von

Ankertests bzw. *Ankeritems*. Ein Ankertest (oder eine Sammlung von Ankeritems) dient generell dazu, die Messergebnisse zweier verschiedener Tests oder Testformen, die dasselbe Merkmal erfassen sollen, so zu adjustieren, dass die resultierenden Testwerte direkt miteinander vergleichbar sind. Mit anderen Worten, die Prozedur der Verankerung stellt Äquivalenz von Tests sicher. Personen, die eine hohe Fähigkeit besitzen, sollten nach Durchführung einer Verankerung bei einem relativ schwierigen Test die gleiche Einstufung ihrer Fähigkeit erfahren wie bei einem relativ leichten Test.

Im Rahmen der Analyse von TestDaF-Erprobungsdaten werden drei verschiedene Rasch-Modelle angewendet. Abbildung 2 gibt eine Übersicht über die Zuordnung der Rasch-Modelle zu den vier Subtests des TestDaF und zum C-Test (die einzelnen Tests sind im linken Teil der Abb. in der Reihenfolge ihrer Darbietung bei einer Erprobungsprüfung aufgeführt).

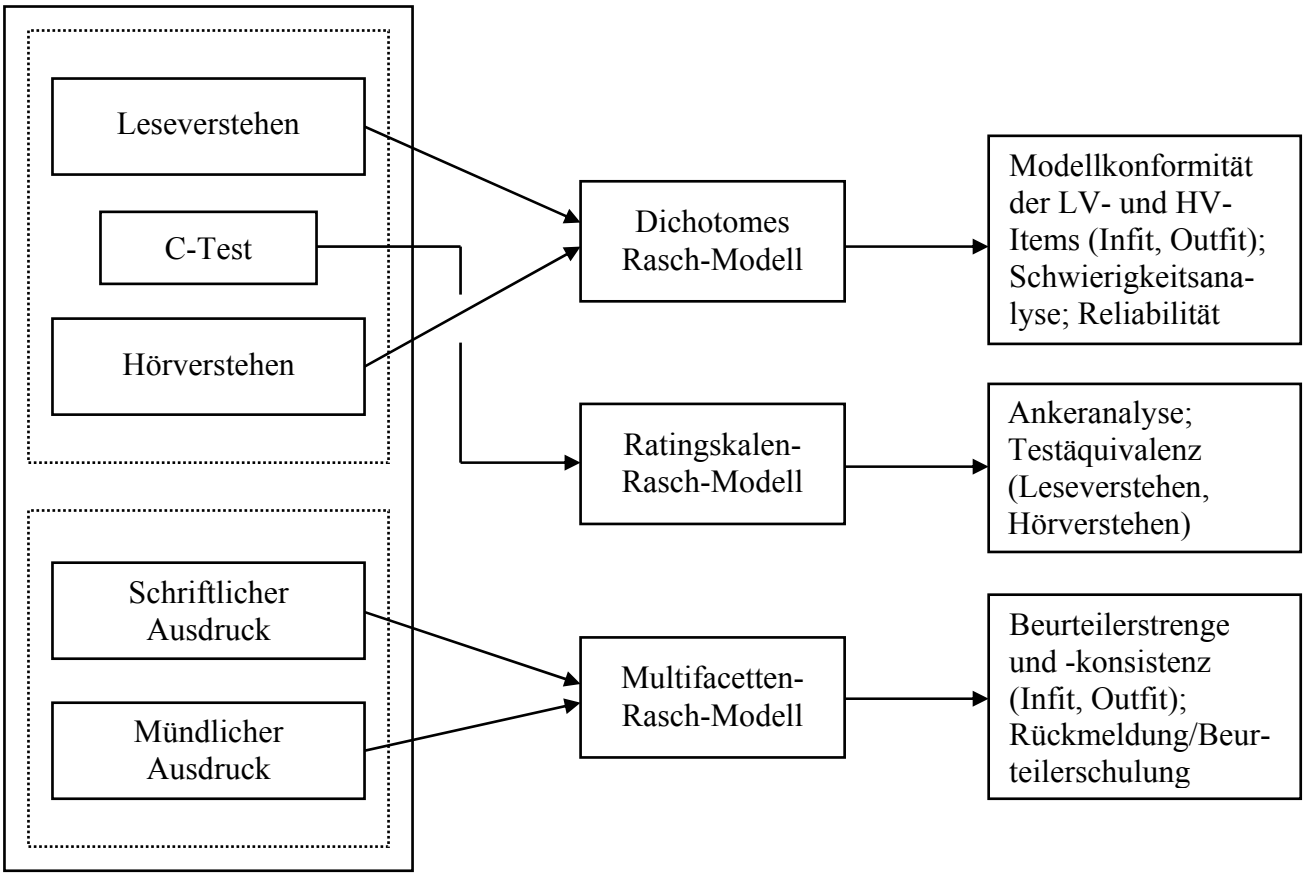

Abbildung 2
Rasch-Modelle zur Qualitätssicherung im Rahmen von TestDaF-Erprobungsprüfungen

3.2 Das dichotome Rasch-Modell

Die Subtests Leseverstehen und Hörverstehen enthalten ausschließlich als „richtig“ oder „falsch“ kodierte Items. Daher wird zur simultanen Skalierung der Personenfähigkeiten und Itemschwierigkeiten das *dichotome Rasch-Modell* verwendet. Hierbei wird die Wahrscheinlichkeit, dass eine Person v Item i richtig beantwortet, d.h. $p(x_{vi} = 1)$, wie folgt ausgedrückt:

$$p(x_{vi} = 1) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}.$$

Die verschiedenen Symbole haben folgende Bedeutung: x_{vi} = Antwort von Person v auf Item i, θ_v = Fähigkeit von Person v (Personenparameter), β_i = Schwierigkeit von Item i (Schwierigkeitsparameter); „exp“ bezeichnet die Exponentialfunktion.

Danach hängt die Wahrscheinlichkeit, dass eine Person v ein Item i richtig beantwortet, nur von der Differenz aus der Fähigkeit θ_v der Person und der Schwierigkeit β_i des Items ab. Ist die Fähigkeit θ_v genauso groß wie die Schwierigkeit β_i, dann resultiert eine Lösungswahrscheinlichkeit von 0.5; ist die Personenfähigkeit größer als die Itemschwierigkeit, dann ist die Wahrscheinlichkeit einer Lösung des Items größer 0.5, im umgekehrten Fall kleiner als 0.5.

Wird die Wahrscheinlichkeit einer richtigen Antwort durch ihre Gegenwahrscheinlichkeit dividiert, erhält man den sog. Wettquotienten (auch „Odds-Ratio“ genannt):

$$\frac{p(x_{vi} = 1)}{p(x_{vi} = 0)} = \exp(\theta_v - \beta_i).$$

Logarithmieren des Wettquotienten ergibt

$$\ln\left[\frac{p(x_{vi} = 1)}{p(x_{vi} = 0)}\right] = \theta_v - \beta_i.$$

Der logarithmierte Wettquotient wird *Logit* genannt („ln“ steht für den natürlichen Logarithmus). Logits sind eine lineare Funktion der Personenfähigkeit θ_v und der Itemschwierigkeit β_i.

3.3 Eine exemplarische Rasch-Analyse zum Leseverstehen

In der logarithmischen Darstellung des dichotomen Rasch-Modells wird deutlich, dass die

beiden Modellparameter θ_v und β_i auf einer gemeinsamen linearen Skala (d.h. auf der *Logitskala*) kalibriert werden. Dies ermöglicht eine anschauliche grafische Darstellung der Personenfähigkeiten und Itemschwierigkeiten in ein und demselben Bezugssystem (vgl. auch McNamara, 1996, pp. 166–168). Personen und Items bilden dabei zwei *Facetten* der Testsituation. Abbildung 3 gibt eine solche Grafik, hier *Facettenraum* genannt, für den Subtest Leseverstehen anhand von Daten aus einer TestDaF-Erprobungsprüfung wieder[5].

Logit	Pbn	Items
	stark	*schwer*
4		
	.	
3	***.	
	***	
	***	
2		
	**.	LV_19
	*****.	LV_16
	********.	
	****	
1	****	LV_18 LV_20
	****.	
		LV_17 LV_21 LV_23
	********	LV_12
	****.	LV_22 LV_26 LV_28
	*********.	LV_10 LV_24
		LV_05
	*****	LV_14
0	********.	LV_02
	*********	LV_25 LV_27
	****	LV_11 LV_15 LV_30
	********	
	****	LV_09
	****	LV_01 LV_08
		LV_07
-1	**.	
	**.	LV_13
		LV_29
	**	LV_03
	*	LV_04 LV_06
	.	
-2		
	.	
-3		
	schwach	*leicht*

Abbildung 3
Exemplarischer Facettenraum für den Subtest Leseverstehen

[5] An dieser Erprobung nahmen 226 Pbn an 4 inländischen und 7 ausländischen Testzentren teil. Die Datenanalysen wurden mit dem Programm WINSTEPS (Version 3.37; Linacre, 2002a) vorgenommen.

Eine solche Darstellung lässt erkennen, wie gut die Fähigkeitsmaße der Personen den Schwierigkeitsmaßen der Items entsprechen. Wenn, wie im vorliegenden Fall, die Verteilung der Personenfähigkeiten hinsichtlich Streuung und Lage entlang der Logitskala mit der Verteilung der Itemschwierigkeiten gut übereinstimmt, dann ist der betreffende Test nicht zu schwer und auch nicht zu leicht für die betrachtete Personenstichprobe. Da leistungsstarke Personen bzw. schwere Items (also z.B. Items 19 und 16) im Facettenraum weiter oben dargestellt werden, ist bei eher leichten Tests die Verteilung der Fähigkeitsmaße relativ zur Verteilung der Schwierigkeitsmaße nach oben verschoben; bei eher schweren Tests verhält es sich genau umgekehrt. Bei geringer Überlappung der beiden Logit-Verteilungen ist der Test nicht ausreichend in der Lage, zwischen Personen unterschiedlicher Fähigkeit über das gesamte Leistungsspektrum hinweg zu differenzieren.

3.4 Modellgeltungskontrolle

Eine IRT-Analyse liefert zu jeder Person und zu jedem Item einen Messwert (d.h. einen Wert auf der Logitskala), 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). Um die Geltung in allen Teilen des Messmodells zu kontrollieren, werden die Abweichungen von den Erwartungen des Modells über alle Personen und alle Items zusammengefasst. In der Regel geschieht dies mittels zweier *Mean-Square-Fehlerstatistiken* (Wright & Masters, 1982, p. 99); diese Statistiken werden auch *Infit*- bzw. *Outfit-Index* genannt (vgl. Eckes, in Druck-a).

Beide Statistiken haben einen Erwartungswert von 1; sie können Werte im Bereich zwischen 0 und $+\infty$ annehmen (Linacre, 2003; Wright & Masters, 1982). Werte deutlich größer 1 verweisen darauf, dass die Daten anhand des Modells nicht gut vorhersagbar sind bzw. mehr Variation aufweisen, als es den Erwartungen des Modells entspricht. Umgekehrt indizieren Werte deutlich kleiner 1, dass ein relativ hohes Maß an Vorhersagbarkeit oder Redundanz vorliegt bzw. die Daten weniger Variation zeigen als vorhergesagt.

Linacre (2002b) hat grobe Richtwerte für die Interpretation von Mean-Square-Statistiken vorgeschlagen. Danach sind Infit- bzw. Outfit-Werte im Intervall zwischen 0.5 und 1.5

messmethodisch akzeptabel[6]. Liegen die Werte der Fit-Statistiken deutlich außerhalb dieser Intervallgrenzen, dann kann dies auch darauf zurückzuführen sein, dass der Test mehr als nur eine Dimension erfasst. Da die verwendeten Rasch-Modelle *eindimensionale* Modelle sind, sie also die beobachteten Antworten auf nur eine einzige Dimension zurückführen, ist in der Residuenanalyse zugleich eine Prüfung der Eindimensionalität eines gegebenen TestDaF-Subtests zu sehen.

Als Maß der Genauigkeit, mit der ein Test als Ganzes die Fähigkeit der Personen unter Anwendung des dichotomen Rasch-Modells misst, dient die *Reliabilität* (auch *Testreliabilität der Personen-Separation* genannt; vgl. Wright & Masters, 1982, p. 106). Diese Form der Reliabilität ist definiert als der Anteil der beobachteten Stichprobenvarianz, der nicht auf Messfehler zurückgeht. Je geringer die Fehlereinflüsse auf die Ergebnisse der Messungen sind, umso höher fällt die Reliabilität aus (das Maximum beträgt 1, das Minimum 0). Im vorliegenden Beispiel belief sich die Reliabilität des Subtests Leseverstehen auf .82, d.h. der Anteil der messfehlerkorrigierten („wahren“) Varianz an der beobachteten Varianz fiel zufriedenstellend hoch aus.

3.5 Ankeranalyse

Wie bereits erwähnt, umfasst die Qualitätssicherung beim TestDaF auch die Sicherstellung gleich bleibender Schwierigkeitsniveaus über verschiedene TestDaF-Prüfungen hinweg. Die Zuordnung der erbrachten Prüfungsleistungen zu Kompetenzstufen setzt notwendig voraus, dass konstruktionsbedingte Schwankungen in der Schwierigkeit der jeweils verwendeten Testsätze keinen nennenswerten Einfluss auf die abschließende Feststellung der Sprachfähigkeit nehmen.

Um der Forderung nach äquivalenten Testsätzen des TestDaF nachzukommen, werden *Ankeranalysen* durchgeführt. Dabei werden die Schwierigkeiten der Items aus den Subtests Leseverstehen und Hörverstehen, wie sie nach einer Erprobung vorliegen, jeweils gemäß der Methode des „Common-Item-Equating“ (Henning, 1987; Wright & Stone, 1979) „verankert“,

[6] Je nach Fragestellung oder Verwendungszusammenhang der Untersuchungsergebnisse können die Intervalle auch breiter oder enger definiert werden (vgl. Bond & Fox, 2001, pp. 176–179).

d.h. zusammen mit anderen, eigens erprobten Items *bekannter* Schwierigkeit auf der gemeinsamen Logitskala kalibriert. Als „gemeinsame“ oder „verbindende“ Items fungieren dabei die Texte eines C-Tests, den die Teilnehmer während einer Erprobungsprüfung im Anschluss an den Subtest Leseverstehen zu bearbeiten haben. Dieser C-Test besteht aus 4 Texten mit jeweils 20 Lücken. Für jeden Text stehen 5 Minuten Bearbeitungszeit zur Verfügung.

In umfangreichen Untersuchungen hat sich der C-Test in seiner Funktion als Ankertest im Rahmen von TestDaF als hervorragend geeignet erwiesen (Arras, Eckes & Grotjahn, 2002; Eckes & Grotjahn, in Druck). Die C-Test-Texte zeichnen sich dadurch aus, dass sie (a) eine hinreichende Differenzierung der Personen hinsichtlich der zu messenden Fähigkeit erlauben, (b) die intendierte Fähigkeit und nur diese messen, (c) Schwierigkeitsmaße aufweisen, die sich über verschiedene Testanwendungen hinweg nicht oder nur unwesentlich verändern, (d) eine ökonomische Durchführung ermöglichen und (e) eine objektive Auswertung gewährleisten.

Die Schwierigkeiten der C-Test-Texte lassen sich nach dem *(diskreten) Ratingskalen-Modell* (Andrich, 1978) oder auch nach dem *kontinuierlichen Ratingskalen-Modell* (Müller, 1999) ermitteln (vgl. Eckes, in Druck-c). Um das Problem der lokalen Abhängigkeit der Lückenwörter innerhalb eines Textes zu lösen, wird der Text als Einheit, d.h. als eine Art *Superitem* oder *Testlet* (mit Itemwerten zwischen 0 und 20) aufgefasst und entsprechend analysiert.

Die testmethodische Qualitätskontrolle bei den produktiven Teilprüfungen Schriftlicher Ausdruck und Mündlicher Ausdruck unterscheidet sich deutlich vom Vorgehen im Falle des Leseverstehens bzw. Hörverstehens. Beim Schriftlichen und Mündlichen Ausdruck nehmen *Beurteiler* eine Einschätzung der Fähigkeit auf der Grundlage der beobachteten Leistungen vor. Im Mittelpunkt der Qualitätssicherung müssen daher die Frage der Fehleranfälligkeit von Leistungsbeurteilungen und die damit verbundene Frage der Beurteilerübereinstimmung stehen (vgl. McNamara, 1996, 2000). Wie beim TestDaF diese Problematik behandelt wird, beschreibt der nächste Abschnitt.

4. Qualitätssicherung bei TestDaF-Leistungsbeurteilungen

4.1 Das Problem mangelnder Beurteilerübereinstimmung

Beurteilungsverfahren zur Abschätzung der Sprachfähigkeit oder zur Messung des Sprachstandes im Rahmen von standardisierten Tests haben in den letzten ca. 15 Jahren zunehmend an Bedeutung gewonnen (vgl. Khattri & Sweet, 1996). Trotz ihrer großen Beliebtheit und weiten Verbreitung sind Beurteilungs- oder Ratingverfahren zur Leistungsmessung grundsätzlich mit einer Reihe von *Urteilsfehlern* behaftet (vgl. z.B. Bortz & Döring, 2002; Hoyt, 2000; Saal, Downey & Lahey, 1980). Diese führen in vielen Fällen zu einer unzureichenden Übereinstimmung zwischen den Beurteilern.

Niedrige Übereinstimmungsraten 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 Teilnehmern) 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 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, in Druck-b; Engelhard, 2002; McNamara, 1996).

4.2 Das Multifacetten-Rasch-Modell

Auf der Grundlage der Item-Response-Theorie sind vielversprechende Ansätze zur Behandlung des Problems geringer Beurteilerübereinstimmung entwickelt worden. Mit der *Multifacetten-Rasch-Analyse* („many-facet Rasch measurement"; kurz MFRM- oder auch *Facets*-Modell; Linacre, 1989; Linacre & Wright, 2002) soll im Folgenden ein spezielles IRT-Modell kurz vorgestellt werden, das es erlaubt, die Beurteilerstrenge zu messen und bei der Festlegung der TestDaF-Niveaustufen zu berücksichtigen (vgl. für eine detailliertere Darstellung Eckes, in Druck-a).

Folgende (systematische) Faktoren bestimmen im typischen Fall die Leistungsbeurteilung: (a) die *Fähigkeit der Personen* (leistungsstärkere Personen sollten höhere Einstufungen, leistungsschwächere Personen niedrigere Einstufungen erhalten), (b) die *Schwierigkeit der Kriterien* (im Schriftlichen Ausdruck; ein „schwieriges" Kriterium ist ein solches, bei dem die Personen generell eher niedrigere Einstufungen erhalten) bzw. die *Schwierigkeit der Aufgaben* (im Mündlichen Ausdruck; eine „schwierige" Aufgabe ist entsprechend eine solche, bei der die Personen generell eher niedrigere Einstufungen erhalten) und (c) die *Strenge der Beurteiler* („strenge" Beurteiler vergeben generell eher niedrigere Bewertungen, „milde" Beurteiler generell eher höhere Bewertungen). Personen, Kriterien bzw. Aufgaben und Beurteiler bilden Facetten der Testsituation.

Allgemeines Ziel einer Multifacetten-Rasch-Analyse ist es, möglichst objektive und präzise Informationen über die Elemente der betrachteten Facetten zu gewinnen. Sie soll also Aufschluss geben nicht nur über die Leistungsfähigkeit der beurteilten Personen, sondern auch über die Schwierigkeit der Aufgaben bzw. Kriterien und die Strenge der Beurteiler. Das hier zugrunde gelegte Multifacetten-Rasch-Modell hat in logarithmischer Schreibweise die folgende 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 Aufgabe i durch Beurteiler j in Kategorie k
- p_{vijk-1} = Wahrscheinlichkeit einer Einstufung von Person v bei Aufgabe i durch Beurteiler j in Kategorie k – 1

- θ_v = Fähigkeitsparameter von Person v
- β_i = Schwierigkeitsparameter von Aufgabe i
- α_j = Strengeparameter von Beurteiler j
- τ_k = Schwierigkeitsparameter von Kategorie k.

In diesem Modell ist der Logit eine lineare Funktion der Personenfähigkeit θ_v, der Aufgabenschwierigkeit β_i, der Beurteilerstrenge α_j und der Kategorieschwierigkeit τ_k. Der Schwierigkeitsparameter von Kategorie (bzw. TestDaF-Niveaustufe) 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 nach dem Ratingskalen-Modell von Andrich, 1978).

4.3 Kontrolle der Beurteilerstrenge

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, sind die beobachteten Einzelurteile in der Weise zu korrigieren, dass leistungs- bzw. konstruktirrelevante Faktoren keinen substanziellen Einfluss auf die endgültige Einstufung haben (Messick, 1989).

Die Beurteilerstrenge ist ein solcher Faktor. Um den Strengefaktor soweit wie möglich zu kontrollieren und um 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 zustande käme, wenn die Person von einem Beurteiler mit durchschnittlicher Strenge beurteilt worden wäre. Dieser hypothetische Beurteiler wäre also weder milder noch strenger als die übrigen Beurteiler. Unterbliebe eine solche Korrektur, würden *leistungsirrelevante* Aspekte der Testsituation (wie die Strenge der Beurteiler) die Messung der Sprachfähigkeit beeinflussen. Bestehen innerhalb der Gruppe von Beurteilern große Unterschiede in der Tendenz zur Strenge bzw. Milde, so würden jene Personen benachteiligt, denen zwei strenge Beurteiler zugeteilt wurden, denn ihre tatsächliche Fähigkeit würde unterschätzt; andere Personen wiederum könnten buchstäblich von Glück reden, wenn ihnen

zwei milde Beurteiler zugewiesen wurden. Neigt der eine Beurteiler zu großer Strenge, der andere aber zu großer Milde, dann wäre eine ausgeprägte Diskrepanz in den Bewertungen derselben Prüfungsleistungen zu erwarten.

Qualitätssicherung beim TestDaF schließt Fairness in der Leistungsbeurteilung ein. Eine *Facets*-Analyse 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 (vgl. Eckes, in Druck-a, in Druck-b). 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 bzw. der jeweiligen Aufgabe bereinigte mittlere Einstufung in der Metrik der Ratingskala an. Auf diese strengekorrigierten, fairen Werte stützt sich die Zuweisung von Niveaustufen bei TestDaF-Prüfungen.

4.4 Eine exemplarische *Facets*-Analyse zum Mündlichen Ausdruck

Die Ergebnisse einer Multifacetten-Rasch-Analyse lassen sich grafisch in Form des Facettenraums darstellen. Analog zur weiter oben beschriebenen Skalierung der Personenfähigkeiten und Itemschwierigkeiten mittels des dichotomen Rasch-Modells (vgl. Abb. 3) erlaubt der Facettenraum im vorliegenden Fall direkte Vergleiche zwischen den Maßen für Personen, Beurteiler und Aufgaben. Abbildung 4 zeigt den Facettenraum im Falle des Mündlichen Ausdrucks, wie er sich für die TestDaF-Prüfung vom April 2003 ergeben hat.[7]

[7] An dieser Prüfung nahmen insgesamt 1.384 Personen teil. Alle hier berichteten Multifacetten-Rasch-Analysen wurden mit dem Programm FACETS (Version 3.40; Linacre, 1999) durchgeführt.

Logit	Pbn	Beurteiler	Aufgaben	TDN
	stark	*streng*	*schwer*	
10	.			(5)
	.			
9	.			
	*.			
8	***.			
	.			
7	****.			
	***.			
6	*.			
	****.			
5	**.			
	**.			-----
4	****.			
	**.		7	
3	****.			
	***.	*		
2	***.		5	4
	***.	**	10	
1	***.	***		
	**.	****	6	
0	**.	******	4 9	
	**.	*******		-----
-1	*.	**		
	*.	*		
-2	*.	**		3
	*.		2 3 8	
-3	.			
	.			
-4	.			-----
	.			
-5	.			
	.			
-6	.			
-7	.			(2)
	schwach	*milde*	*leicht*	

Abbildung 4
Exemplarischer Facettenraum für den Subtest Mündlicher Ausdruck

Der Facettenraum ist in fünf Spalten unterteilt. Die erste Spalte enthält die Logitskala. Ein Wert auf dieser Skala gibt das Fähigkeitsmaß einer Person, das Strengemaß eines Beurteilers oder das Schwierigkeitsmaß einer Aufgabe wieder. Die zweite Spalte zeigt die Verteilung der Parameterschätzungen im Hinblick auf die sprachliche Leistungsfähigkeit der Pbn Leistungsstärkere Pbn sind im Facettenraum weiter oben, leistungsschwächere Pbn weiter unten abgebildet (jedes Sternchen steht für 20 Pbn, jeder Punkt für 1 bis 19 Pbn). 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 (jedes Sternchen steht hier für 1 Beurteiler). Ganz offenkundig ist die Variabilität innerhalb der Beurteilerfacette substanziell: Die Strengemaße reichen von 2.45 am oberen Ende der Logitskala bis –2.22 am unteren Ende (hierauf wird später noch genauer eingegangen). In der vierten Spalte sind die neun Aufgaben des Mündlichen Ausdrucks entsprechend ihrer Schwierigkeitsmaße

angeordnet.[8] Die fünfte und letzte Spalte bildet die TDN-Skala auf die Logitskala ab, d.h. sie gibt die Fähigkeit der Personen in der Metrik der TDN-Skala wieder.

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 *Facets*-Modells in Einklang steht. Da dieses Modell die Leistungsbeurteilung als einen stochastischen Prozess konzipiert, wird ein gewisses Maß an zufälliger Schwankung modellimmanent vorausgesetzt.

Auskunft über den Grad der Konsistenz jedes einzelnen Beurteilers geben die Werte der Infit- und Outfit-Statistiken. Diese Statistiken fassen im vorliegenden Anwendungsfall das Ausmaß der Schwankungen über alle jeweils beurteilten Pbn und über alle Aufgaben zusammen (vgl. Eckes, in Druck-a). Sie zeigen an, inwieweit die Bewertungen eines gegebenen Beurteilers mehr oder weniger Variation aufweisen, als vom Modell erwartet wird. 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.

Die ermittelten Konsistenz- und Strengewerte können im Rahmen einer *Rückmeldung* an die Beurteiler weitergegeben werden. Diese Informationen helfen, Fragen der Beurteiler bezüglich der eigenen Bewertungsleistung zu beantworten, eventuell vorhandene Tendenzen bei den Bewertungen aufzudecken und mögliche Ursachen hierfür zu diskutieren. Schließlich trägt eine derartige Rückmeldung dazu bei, die Qualität des gesamten Korrekturprozesses zu erhöhen bzw. auf Dauer zu sichern.

Wie bereits ausgeführt, hat sich der Strengeeffekt in der Forschung zu Sprachprüfungen als ein sehr starker und robuster Effekt erwiesen. Ziel von *Beurteilertrainings* sollte es daher nicht sein, das Korrekturverhalten dahingehend zu verändern, dass alle Beurteiler denselben Bewertungsstandard verwenden. Vielmehr sollte es darum gehen, die Bedeutung und

---

[8] Die Aufgabenschwierigkeiten entsprechen nur bedingt der angezielten Anordnung (Aufgaben 5, 7 und 10 sollten auf TDN 5-Niveau, Aufgaben 4, 6 und 9 auf TDN 4-Niveau sowie Aufgaben 2, 3 und 8 auf TDN 3-Niveau liegen). So könnten die Aufgaben aus der TDN 5-Gruppe noch etwas schwieriger sein, um besser zwischen leistungsstarken Pbn zu differenzieren. Zudem unterscheiden sich einzelne Aufgaben innerhalb der TDN 5- bzw. der TDN 4-Gruppe in ihren Schwierigkeitsmaßen signifikant voneinander. Gegenwärtig wird das Format des Mündlichen Ausdrucks überarbeitet, um auch an diesem Punkt Verbesserungen zu erzielen.

Anwendung der jeweiligen Beurteilungskriterien zu klären und den korrekten Gebrauch der TDN-Skala zu erläutern. Die Anstrengungen sollten sich also auf die Erhöhung bzw. Stabilisierung der Konsistenz *innerhalb* der Beurteiler und nicht auf die Beseitigung der Strengeunterschiede *zwischen* den Beurteilern richten.

4.5 Strenge und Konsistenz in verschiedenen TestDaF-Prüfungen

Einen Überblick über die Ergebnisse der *Facets*-Analysen von Leistungsbeurteilungen in verschiedenen TestDaF-Prüfungen geben die nachfolgenden Tabellen. Die Ergebnisse zum Schriftlichen Ausdruck finden sich in Tabelle 1, die Ergebnisse zum Mündlichen Ausdruck in Tabelle 2. Dabei ist zu beachten, dass diese Tabellen nur einen sehr kleinen Ausschnitt aus der Fülle an Informationen wiedergeben, die eine *Facets*-Analyse für Zwecke der Qualitätssicherung beim Sprachtesten bereitstellt. Die folgende Darstellung beschränkt sich aus Raumgründen ganz auf ausgewählte Ergebnisse zur Beurteilerstrenge und Beurteilerkonsistenz.

Tabelle 1.
Strenge und Konsistenz von Beurteilern in verschiedenen TestDaF-Prüfungen – Schriftlicher Ausdruck

Statistik	T003	T004	T005	T006	T007	T008
Anzahl der Beurteiler	29	28	32	21	21	30
Strenge (max / min)[a]	2.17 / –2.08	3.31 / –6.64	2.18 / –2.92	4.66 / –3.58	2.07 / –3.37	4.08 / –4.06
Homogenitätstest[b]	1828.5*	2371.6*	2052.9*	1682.3*	971.6*	4803.6*
Klassenseparation[c]	9.59	12.25	10.24	11.49	8.88	17.40
Separationsreliabilität[d]	0.98	0.99	0.98	0.99	0.98	0.99
Konsistenzindex A[e] $0.5 \leq$ Infit ≤ 1.5	29	28	32	21	21	30
Konsistenzindex B[e] $0.5 \leq$ Outfit ≤ 1.5	29	27	32	19	21	29

Anmerkung: [a] Logitwerte. [b] Chi-Quadrat-Test mit $df = J - 1$, wobei J = Anzahl der Beurteiler. [c] Anzahl statistisch reliabel unterscheidbarer Klassen von Beurteilern. [d] Genauigkeit, mit der die Strengewerte voneinander unterschieden werden können. [e] Angegeben ist die Anzahl der Beurteiler, deren Konsistenzwerte in das Akzeptanzintervall fallen. * $p < .01$.

Tabelle 2.
Strenge und Konsistenz von Beurteilern in verschiedenen TestDaF-Prüfungen – Mündlicher Ausdruck

Statistik	T003	T004	T005	T006	T007	T008
Anzahl der Beurteiler	31	32	36	30	28	35
Strenge (max / min)[a]	1.37 /–1.47	2.57 / –2.80	1.84 / –1.83	2.53 / –3.50	2.45 / –2.22	2.65 / –2.61
Homogenitätstest[b]	2174.2*	2656.3*	2789.2*	2309.1*	1608.7*	2768.9*
Klassenseparation[c]	10.85	12.00	10.15	11.00	10.07	10.79
Separationsreliabilität[d]	0.98	0.99	0.98	0.98	0.98	0.98
Konsistenzindex A[e] $0.5 \leq \text{Infit} \leq 1.5$	31	32	36	28	27	35
Konsistenzindex B[e] $0.5 \leq \text{Outfit} \leq 1.5$	28	32	35	27	23	33

Anmerkung: [a] Logitwerte. [b] Chi-Quadrat-Test mit $df = J - 1$, wobei J = Anzahl der Beurteiler. [c] Anzahl statistisch reliabel unterscheidbarer Klassen von Beurteilern. [d] Genauigkeit, mit der die Strengewerte voneinander unterschieden werden können. [e] Angegeben ist die Anzahl der Beurteiler, deren Konsistenzwerte in das Akzeptanzintervall fallen. * $p < .01$.

Die Maxima und Minima der Logitwerte für den Strengeparameter geben einen ersten Hinweis darauf, wie unterschiedlich die Beurteiler innerhalb der jeweiligen Gruppen sind. Diese Unterschiedlichkeit erweist sich im Homogenitätstest in allen Fällen als statistisch hochsignifikant, d.h. die Annahme homogener Strengemaße ist eindeutig zurückzuweisen. Auch der Index der Klassenseparation zeichnet ein klares Bild: Die Zahl der statistisch reliabel unterscheidbaren Klassen von Beurteilern liegt in beiden Subtests bei etwa 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 (im Falle homogener Strengemaße ginge diese Reliabilität gegen 0).

Hinsichtlich der Beurteilerkonsistenz enthalten die Tabellen Angaben darüber, wie viele Beurteiler in einer gegebenen Prüfung Infit- bzw. Outfit-Werte außerhalb der Grenzen des 0.5/1.5-Intervalls aufweisen. Was den Schriftlichen Ausdruck betrifft, so liegen die Infit-Werte

für alle Beurteiler im akzeptablen Bereich; bei den Outfit-Werten gibt es nur ganz vereinzelt Hinweise auf Inkonsistenz. Bezüglich des Mündlichen Ausdrucks fallen die Konsistenzergebnisse sehr ähnlich aus; die wenigen Inkonsistenzen betreffen fast ausnahmslos den Outfit-Index. Dabei ist zu berücksichtigen, dass der Outfit-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. Insgesamt kann also von einer außerordentlich hohen Konsistenz der Beurteiler in allen hier betrachteten TestDaF-Prüfungen gesprochen werden.

5. Zusammenfassung und Diskussion

Angesichts der stark gestiegenen gesellschaftlichen Bedeutung von Sprachprüfungen ist eine sorgfältige Kontrolle und Sicherung ihrer Qualität mehr denn je geboten. Eine zentrale Rolle kommt hierbei der konsequenten Anwendung methodischer Standards der Testkonstruktion und Testevaluation zu. Entsprechende Qualitätskriterien finden aber gerade im Kontext des Sprachtestens immer noch viel zu wenig Berücksichtigung. Dies ist umso problematischer, als Sprachtestergebnisse häufig weitreichende Folgen für die Testteilnehmer haben.

Im vorliegenden Beitrag wurden am Beispiel der weltweit durchgeführten Sprachprüfung TestDaF einige Aspekte einer testmethodisch fundierten Qualitätssicherung aufgezeigt. Als Kernstück wurde der Erprobungszyklus vorgestellt. Mit seinen drei eng aufeinander bezogenen Hauptphasen der Vorerprobung, Revision und Erprobung zielt dieser Zyklus auf eine Optimierung der Qualität aller vier Teilprüfungen des TestDaF.

Qualitätssicherung als dynamischer zielgerichteter Prozess erfordert die ständige Bereitschaft, einzelne Testelemente wie auch das Design einzelner Subtests auf der Basis der testmethodisch gewonnenen Erkenntnisse neu zu überdenken und ggf. zu überarbeiten. Erfüllen die modifizierten Komponenten des Tests die angezielten Qualitätskriterien, so kann eine Anwendung bei realen Testadministrationen verantwortet werden.

Der Schwerpunkt der testmethodischen Analysen von TestDaF-Daten liegt auf Modellen, die sich aus der Item-Response-Theorie ableiten. Erst die verschiedenen Rasch-Modelle, die beim

TestDaF routinemäßig zum Einsatz kommen, können mit hinreichender Genauigkeit Fehler und Schwächen in neu erstellten Aufgabensammlungen identifizieren und diese einer fundierten Revision zuführen (Embretson & Reise, 2000; Hambleton et al., 2000; Stone, 2002). Die hohe Flexibilität von IRT-Modellen ist es auch, die sie zu einem besonders geeigneten Instrument zur Herstellung von Testäquivalenz macht. Äquivalenz unterschiedlicher TestDaF-Prüfungen stellt eine stabile, der Fähigkeit der Pbn entsprechende Zuordnung zu TestDaF-Niveaustufen sicher.

Eine weit über den Kontext von Sprachprüfungen hinausreichende Bedeutung hat die Kontrolle des Strengefaktors bei Leistungsbeurteilungen im Rahmen des Multifacetten-Rasch-Modells (Eckes, in Druck-a; Linacre & Wright, 2002). Ob bei der Notengebung in der Schule oder bei der Mitarbeiterbeurteilung im Betrieb, Beurteiler stellen eine einflussreiche Quelle konstruktirrelevanter Varianz dar, die soweit wie möglich einzudämmen ist. Wichtige Impulse hierzu können von Erkenntnissen ausgehen, die mit dem Multifacetten-Rasch-Modell beim Sprachtesten gewonnen werden.

Literatur

American Educational Research Association (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

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

Arras, U., Eckes, T. & Grotjahn, R. (2002). C-Tests im Rahmen des „Test Deutsch als Fremdsprache“ (TestDaF): Erste Forschungsergebnisse. In R. Grotjahn (Hrsg.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen*. Bd. 4. Bochum: AKS-Verlag, 175–209.

Arras, U. & Grotjahn, R. (2002). TestDaF: Aktuelle Entwicklungen. *Fremdsprachen und Hochschule*, *66*, 65–88.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*, 1–42.

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

Bortz, J. & Döring, N. (³2002). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Berlin: Springer-Verlag.

Eckes, T. (in Druck-a). Beurteilerübereinstimmung und Beurteilerstrenge: Eine Multifacetten-Rasch-Analyse von Leistungsbeurteilungen im „Test Deutsch als Fremdsprache“ (TestDaF). *Diagnostica*.

Eckes, T. (in Druck-b). Facetten des Sprachtestens: Strenge und Konsistenz in der Beurteilung sprachlicher Leistungen. In A. Wolff et al. (Hrsg.), *Materialien Deutsch als Fremdsprache*. Regensburg: FaDaF.

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

Eckes, T. & Grotjahn, R. (2003, July). *C-tests as measures of general language proficiency*. Paper presented at the 25th Language Testing Research Colloquium, Reading, UK.

Eckes, T. & Grotjahn, R. (in Druck). Der C-Test als Ankertest für TestDaF: Analysen auf der Basis eines probabilistischen Testmodells. In R. Grotjahn (Ed.), *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. (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*. Mahwah, NJ: Erlbaum, 261–287.

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.

Fischer, G. H. (1996). IRT-Modelle als Forschungsinstrumente der Differentiellen Psychologie. In K. Pawlik (Hrsg.), *Grundlagen und Methoden der Differentiellen Psychologie*. Göttingen: Hogrefe, 673–729.

Grotjahn, R. (2000). Testtheorie: Grundzüge und Anwendungen in der Praxis. In A. Wolff & H. Tanzer (Hrsg.), *Sprache – Kultur – Politik*. Regensburg: FaDaF, 304–341.

Häcker, H., Leutner, D. & Amelang, M. (Hrsg.) (1998). *Standards für pädagogisches und psychologisches Testen*. Göttingen: Hogrefe.

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*. San Diego, CA: Academic Press, 553–581.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Boston, MA: Heinle & Heinle.

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.

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

Kniffka, G. & Üstünsöz-Beurer, D. (2001). TestDaF: Mündlicher Ausdruck. Zur Entwicklung eines kassettengesteuerten Testformats. *Fremdsprachen Lehren und Lernen*, *30*, 127–149.

Kubinger, K. D. ([4]1999). Testtheorie: Probabilistische Modelle. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik: Ein Lehrbuch*. Weinheim: Psychologie Verlags Union, 322–334.

Lienert, G. A. & Raatz, U. ([6]1998). *Testaufbau und Testanalyse*. 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. (2002a). *A user's guide to WINSTEPS-MINISTEP: Rasch-model computer programs*. Chicago: MESA-Press.

Linacre, J. M. (2002b). 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.

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

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

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

Moosbrugger, H. ([4]1999). Testtheorie: Klassische Ansätze. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik: Ein Lehrbuch*. Weinheim: Psychologie Verlags Union, 310–322.

Müller, H. (1999). *Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen: Einführung in die Item-Response-Theorie für abgestufte und kontinuierliche Items*. Bern: Huber.

Projektgruppe TestDaF (2000). TestDaF: Konzeption, Stand der Entwicklung, Perspektiven. *Zeitschrift für Fremdsprachenforschung*, *11*, 63–82.

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

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion*. 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.

Steyer, R. & Eid, M. (²2001). *Messen und Testen*. Berlin: Springer-Verlag.

Stone, M. H. (2002). Quality control in testing. *Popular Measurement*, *4*, 15–23.

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

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.