

Sprachtests müssen fair sein

Am Beispiel der telc-Prüfungen im Bereich Migration und Einbürgerung

Peter Kiefer

Im Rahmen der gesellschaftlich und politisch eminent wichtigen Themenbereiche Integration und Einbürgerung kam es europaweit in den zurückliegenden fünfzehn Jahren zu einer Vielzahl von Verordnungen und gesetzlichen Regelungen – und entsprechend zu einer unüberschaubaren Flut von Publikationen und Äußerungen in den unterschiedlichsten Medien.

Was die Situation in Deutschland betrifft, so wären in diesem Zusammenhang das seit 2005 geltende Zuwanderungsgesetz und die 2007 von der Innenministerkonferenz der Länder beschlossene Standardisierung des Einbürgerungsverfahrens zu nennen. In beiden Fällen ist das Erreichen einer bestimmten sprachlichen Kompetenz – Nachweis der Sprachkompetenz auf dem Europaratsniveau B1 – ein konstitutives Element. Da die Überprüfung von Sprachkenntnissen der Migranten und Einbürgerungswilligen prinzipiell ein hochsensibler Bereich ist, sollte sie ausschließlich auf der Grundlage fachwissenschaftlicher Expertise und in enger Kooperation mit den entsprechenden Fachinstituten vorgenommen werden. Insofern ist es beruhigend, dass seit Inkrafttreten des Zuwanderergesetzes die Durchführung des sprachlichen Abschlusstests der Integrationskurse in die Hände der telc GmbH als Exklusivpartner des Bundesamts für Migration und Integration gelegt wurde. Die telc GmbH ist Entwickler und Träger des *Zertifikat Deutsch* und zahlreicher weiterer Sprachtests¹, verfügt als Prüfungszentrale und Tochterunternehmen des Deutschen Volkshochschulverbands über Hunderte seriöser Prüfungszentren und blickt zusammen mit ihren Vorort-Partnern, den deutschen Volkshochschulen, auf eine jahrzehntelange Erfahrung bei der Feststellung der sprachlichen Kompetenzen von Migranten und Bewerbern um die deutsche Staatsbürgerschaft zurück. Als Mitglied der ALTE (Association of Language Testers in Europe) unterliegt die telc GmbH der Selbstverpflichtung, die Standards für objektive und faire Testverfahren konstant hochzuhalten. Auf diesem Hintergrund ist es gut und wichtig, dass sich eine Expertengruppe der ALTE bereits seit Jahren des Themas „Sprachtests für Migranten“ angenommen hat und unbedingt erforderliche Mindeststandards für Prüfungen für die genannten Zielgruppen definiert.² Die zentralen Kriterien für die Standards sollen auf diesen

Seiten vorgestellt und diskutiert werden. Dabei werden sowohl das in den oben genannten Verfahren bewährte *Zertifikat Deutsch* (B1) als auch die laufenden Arbeiten des zurzeit von der telc GmbH und dem Goethe-Institut entwickelten *Deutsch-Tests für Zuwanderer* in den Blick genommen.

Neben vielen Einzelkriterien, die für die Qualität eines Sprachtests herangezogen werden, ist vor allem der Aspekt der Fairness zu nennen. Gerade in einem so sensiblen Bereich, in dem Testverfahren rechtliche Konsequenzen (z. B. Aufenthaltserlaubnis) für Prüfungskandidaten haben können, ist Fairness eine *conditio sine qua non*.

Einer der zentralen Bestandteile eines fairen Testverfahrens ist *Transparenz*. Diese ist im *Zertifikat Deutsch* umfänglich und in geradezu vorbildlicher Weise hergestellt: Veröffentlicht ist nicht nur ein feststehendes Testformat mit genauer Definition der Subtests in den klassischen vier Fertigungsbereichen und klar handhabbaren Bewertungskriterien, sondern auch eine detaillierte Lernzielbeschreibung, die die sprachlichen Handlungen, die die Prüfungskandidaten beherrschen sollen, als ganze und in ihren Einzelbestandteilen inventarisiert. Für alle, die sich mit Sprachtests professionell beschäftigen, mag dies eine Selbstverständlichkeit sein, bei der Überprüfung von Deutschkenntnissen im Rahmen der Einbürgerung war die Praxis der vergangenen Jahre in zahlreichen Bundesländern von der geforderten Transparenz jedenfalls weit entfernt. So mutet es im Nachhinein geradezu abenteuerlich an, dass sich Menschen, die sich um die deutsche Staatsbürgerschaft bemühten, nicht einen für alle gleichen standardisierten Sprachtest ablegen mussten wie z. B. das *Zertifikat Deutsch*, sondern mehr oder weniger willkürlich der Einschätzung von fachfremdem und ungeschultem Personal in einem Gespräch über einen aktuellen Zeitungsartikel ausgesetzt waren. So manch ein Mitarbeiter in den Behörden agierte frei nach dem Motto: „Feststellen, ob bzw. wie gut jemand Deutsch kann, das kann doch jeder...“ Eine hohe Meinung über die Arbeit professioneller Sprachtester ist über solche oder ähnlich lautende Statements natürlich nicht zum Ausdruck gebracht worden. Aber die Zeiten ändern sich, und die Wertschätzung der Arbeit fachwissenschaftlich fundiert zu Werke gehender Testentwickler steigt.

Doch zurück zum *Zertifikat Deutsch* als einer Sprachprüfung mit Beispielcharakter: Die Eignung des Zertifikats als Feststellungsprüfung für Einbürgerungswillige und Migranten liegt nicht zuletzt in seiner Ausrichtung auf kommunikative Handlungskompetenzen. Diesbezüglich ist es hilfreich, sich die Wirkmechanismen der Einzelbestandteile des Zertifikats wie das Ineinandergreifen des Räderwerks bei einer komplexen Maschine vorzustellen. Deshalb sei zunächst nochmals daran erinnert, dass das Zertifikat gar nicht nur aus den bloßen Prüfungsaufgaben, sondern aus einem Gesamtkomplex eines standardisierten Aufgabenapparats besteht, der mehrfach mit detaillierten Lernzielpezifikationen vernetzt – man könnte sagen: durch diese abgesichert – ist. Jede Einzelaufgabe des Zertifikats als zu erledigende kommunikative Handlung ist also nur im Rahmen eines präzise definierten curricularen Spielraums auf dem Europaratsniveau B1 möglich, d. h. Fairness und Transparenz sind Maßstäbe für jede einzelne „Task“ (kommunikativer Handlungsauftrag).

Wie funktioniert nun die transparente Begrenzung und Beschreibung von Tasks? Zur besseren Veranschaulichung konstruieren wir ein Beispiel aus dem Bereich der

mündlichen Interaktion: In einem Konsensfindungsgespräch im Paarprüfungsformat des Zertifikats (B1) sollen zwei Kandidaten gemeinsam ein Lernwochenende planen. In einem ersten Schritt stützt sich diese Aufgabe auf diverse niveaubezogene Kann-Beschreibungen des Gemeinsamen europäischen Referenzrahmens des Europarats³.

Der/die Lernende

- kann in Diskussionen die eigenen Ansichten durch relevante Erklärungen, Argumente und Kommentare begründen und verteidigen (B2)
- kann seine/ihre Meinung oder Reaktion klar machen, wenn es darum geht, ein Problem zu lösen oder praktische Fragen zu klären im Zusammenhang damit, wohin man gehen oder was man tun sollte
- kann höflich Überzeugungen und Meinungen, Zustimmung und Ablehnung ausdrücken. (B1)
- kann Vorschläge machen und auf Vorschläge reagieren
- kann anderen beipflichten oder widersprechen
- kann auf einfache Weise praktische Fragen des Alltags besprechen, wenn er/sie klar, langsam und direkt angesprochen wird (A2).⁴

Die hier angeführten Deskriptoren der drei Niveaustufen A2-B1-B2 sind lediglich Beispiele aus einem großen Repertoire an Kann-Beschreibungen und sollen – in unserem Fall – auch zeigen, wie das Zielniveau des Zertifikats (B1) von den benachbarten Stufen abgegrenzt werden kann. So wird schon im Bereich dieser mittels Kann-Beschreibungen definierten „Groblerziele“ deutlich, dass auf der Elementarstufe A2 noch kein Gespräch im eigentlichen Sinn geführt werden kann. In der Terminologie der telc-Zertifikate heißt das, dass auf dem Niveau A2 noch kein „Szenario“ (hier: Konsensfindung) mit diskursiven und strategischen Redemitteln realisiert werden soll, sondern die mündliche Interaktion relativ unverbunden hauptsächlich auf der funktionalen Ebene stattfinden kann. Die Beobachtung von Kandidatengesprächen in der Prüfung Start Deutsch 2 (A2) bestätigt das in jeder Hinsicht: Auf dem Weg hin zum echten Konsensfindungsdiskurs (B1/B2) sind die Lernenden auf dem Niveau A2 vorläufig auf die Realisierungsmöglichkeiten der Sprachintention „Vorschläge machen/auf Vorschläge reagieren“ begrenzt. Nimmt man die Abgrenzung unseres Zielniveaus B1 nach oben in den Blick, so wird deutlich, dass ein Kommentieren und Argumentieren, z. B. auch im Rahmen komplexerer Themen- und Aufgabenstellungen, wie es auf der Stufe B2 verlangt wird, von den Integrationskursteilnehmern auf dem Niveau B1 noch nicht geleistet werden kann. Sehr wohl können sie aber im Rahmen des ihnen vertrauten Themenbereichs „Deutsch lernen“ gemeinsam ein Lernwochenende planen, können ihre Vorschläge, Gegenvorschläge, Meinungen und Überzeugungen ansatzweise strategisch und diskursiv gestalten. Ganz allgemein kann man davon ausgehen, dass alle telc-Deutschtests von Start Deutsch 1 (A1) bis telc Deutsch C1, aber insbesondere auf der Strecke von Start Deutsch 2 (A2) über das *Zertifikat Deutsch* (B1) bis zu telc Deutsch B2 hinsichtlich ihrer Lernzielbeschreibungen und Aufgabentypologien nicht nur eine grammatikalische und lexikalische, sondern auch eine funktionale, strategische und diskursive Progression enthalten.

Bevor wir uns auf einige exemplarisch herauszustellende „Feinlernziele“ konzentrieren, wollen wir den Blick zunächst noch einmal ausweiten auf das Gesamtgeschehen von curricular fundiertem Sprachenlernen und Feststellung von Kompetenzen mittels standardisierter Tests. Das Leitprinzip Handlungsorientierung weist an dieser Stelle nämlich über Curriculum, Unterricht und Test hinaus: Die auf der theoretischen Ebene im Curriculum mittels Kann-Beschreibung definierten sprachlichen Handlungen (wie z. B. Konsensfindungen oder informelle Diskussionen) haben bei rein schematischer Betrachtung drei Bezugspunkte zur kommunikativen Praxis: als Aufgabe im Sprachtraining, als Aufgabe in der Sprachprüfung – und als Aufgabe in der sprachlichen Wirklichkeit des Alltags. Auf diesem Hintergrund wird Sprachkompetenz qua Handlungskompetenz als Voraussetzung für politische Integration und Partizipation bedeutsam. Es zeigt sich einmal mehr, dass Curriculum, Unterricht und Test grundsätzlich als Gesamtkonzeption aufzufassen sind und es zwangsläufig zu Ausfällen kommen muss, wenn an einer Stelle im System dieser kommunizierenden Röhren ein Fehler auftritt. Zur Veranschaulichung diene ein ganz einfaches Beispiel: Wenn die Unterrichtspraxis auf ein Lexik-Grammatik-Training reduziert wird, ist die Wahrscheinlichkeit hoch, dass die Teilnehmenden in einer handlungstheoretisch begründeten Prüfung nicht besonders gut abschneiden. Umgekehrt ist ein kommunikativer Unterricht, der als solcher stets auch die sprachliche Wirklichkeit und die Handlungsbedarfe der Teilnehmenden mit einbezieht, die beste Voraussetzung für ein gutes Prüfungsergebnis im *Zertifikat Deutsch* – denn diese Prüfung beinhaltet hinsichtlich ihrer Lernziele ein „common core“ an kommunikativem Repertoire auf einem angemessenen Kompetenzniveau (B1). Der „common-core“-Aspekt sollte auch bei einer Prüfungskonzeption – wie bei dem zurzeit zu entwickelnden Deutsch-Test für Zuwanderer – nie außer acht gelassen werden, da sich auf den Stufen A2 und B1 eine Ausrichtung auf zu spezielle Wort- oder Handlungsfelder als kontraproduktiv erweisen würde. Typische Messfehler (*bias*) wären beispielsweise zu erwarten, wenn bei einer Überbetonung des Themenbereichs Gesundheit Teilnehmende ohne besondere Erfahrung auf diesem Gebiet schlechtere Ergebnisse erzielen würden als Personen, die z. B. als Krankenpfleger(innen) über Weltwissen verfügen. Eine solche Prüfung wäre für die Mehrzahl der Teilnehmenden unfair.

Im Anschluss daran nun noch ein Wort zu den „Feinlernzielen“ des *Zertifikat Deutsch*; denn eine objektive und faire Bewertung von Teilnehmerleistungen ist natürlich nur möglich, wenn die Kriterien dafür an den verschiedenen Lernzielinventaren ausgerichtet sind. Was heißt das im einzelnen, wenn wir, wiederum exemplarisch, von der Bewertung mündlicher Interaktion sprechen?

Wurde per „Groblernziel“ mittels eines Pakets von Kann-Beschreibungen eine Niveaustufe definiert (z. B. die Stufe B1 in Abgrenzung zu A2 und B2), so geben die inventarisierten „Feinlernziele“ den Prüfenden präzise Anhaltspunkte für ihre Bewertung. Gerade im Bereich der produktiven Fertigkeiten, wo Teilnehmerleistung nicht durch eine IT-gestützte Auswertungsapparatur erfolgt, sondern durch Menschen, ist im Sinne von Fairness und Vergleichbarkeit ein Höchstmaß von Objektivität nur zu gewährleisten in der strikten Anwendung von an „Feinlernziele“ gebundenen Bewertungskriterien. Im *Zertifikat Deutsch* erfolgt die Bewertung jeder mündlichen Teil-

nehmerleistung durch zwei lizenzierte Prüfende, die in einem Prüfertraining, welches alle drei Jahre absolviert werden muss, die faire und objektive Handhabung der Bewertungskriterien einstudieren. Die Vergabe der Prüferlizenz erfolgt nicht aufgrund bloßer Teilnahme an dem Training, sondern durch Erbringen bestimmter Leistungen. Die Arbeit in den Prüfertrainings ist bei telc gerade darauf ausgerichtet, dass Prüfende die auf diesen Seiten beschriebenen Zusammenhänge von Lernzielen, Prüfungsaufgaben, Teilnehmerleistungen und deren Beurteilung im Detail nachvollziehen und verstehen. Die Bewertung der mündlichen Leistungen im *Zertifikat Deutsch* erfolgt bekanntlich anhand von vier Kriterien. Das Kriterium I („Ausdrucksfähigkeit“) bezieht sich u. a. auf die Realisierung aufgabentypischer Sprachintentionen sowie auf das performierte Lexikrepertoire. Prüfende haben demzufolge bei der Bewertung der Ausdrucksfähigkeit die niveauspezifischen Lernzielinventare „Sprachintentionen“ und „Lexik“ mit Blick auf die jeweilige Aufgabe zu Rate zu ziehen. Bei unserem Konsensfindungs-Beispiel geht es beispielsweise um das Realisierungsspektrum der Sprachintention „Vorschläge machen/auf Vorschläge reagieren“ und um einschlägigen Wortschatz zum Thema „Lernwochenende/Deutsch lernen“. Für beides gibt es im *Zertifikat Deutsch* listenartige Aufstellungen. Die Rückbindung der Bewertung z. B. im Rahmen von Kriterium I an die jeweiligen Inventare macht den ganzen Prozess objektiv nachvollziehbar und garantiert bei Einhaltung aller Regeln auch zu mehr Fairness.

Im Bereich der rezeptiven Fertigkeiten, wo telc-Prüfungen über eine IT-gestützte Apparatur ausgewertet werden, sorgt die Einhaltung des von ALTE definierten Produktionsprozesses in der Testentwicklung für Objektivität und Fairness. Prinzipiell gilt natürlich auch hier die Anbindung aller Testitems an das fertikeitsbezogene Lernzielniveau, also beispielsweise rekurriert der Bereich „globales Leseverstehen“ u. a. auf folgende Deskriptoren. Das Zielniveau B1 für das *Zertifikat Deutsch* ist wiederum gegenüber den benachbarten Stufen A2 und B2 typographisch hervorgehoben.

Der/die Lernende

- kann aus hoch spezialisierten Quellen des eigenen Fachgebiets Informationen, Gedanken und Meinungen entnehmen; kann Fachartikel, die über das eigene Gebiet hinausgehen, lesen und verstehen, wenn er/sie ab und zu im Wörterbuch nachschlagen kann, um das Verständnis der verwendeten Terminologie zu überprüfen
- kann in unkomplizierten Zeitungsartikeln zu vertrauten Themen die wesentlichen Punkte erfassen (B2)
- kann in klar geschriebenen argumentativen Texten die wesentlichen Schlussfolgerungen erkennen; kann bei der Behandlung eines Themas die Argumentation erfassen, wenn auch nicht unbedingt im Detail
- kann Artikel und Berichte zu aktuellen Fragen lesen und verstehen, in denen die Schreibenden eine bestimmte Haltung oder einen bestimmten Standpunkt vertreten (B1)
- kann aus einfacheren schriftlichen Materialien wie Briefen, Broschüren oder Zeitungsartikeln, in denen Ereignisse beschrieben werden, spezifische Informationen herausfinden (A2).⁵

Wie bei den produktiven Fertigkeiten reichen die Kann-Beschreibungen als „Groblernziel“ nicht aus für die Konstruktion einer niveaugerechten Aufgabe. Eine ganze Reihe von Prozeduren in der Testentwicklung sind einzuhalten, um z. B. ein Leseverstehen-Item für den Echteininsatz in einem Test für gut zu befinden. Der Weg muss in diesem Fall immer über Erprobungen und statistische Fachanalysen führen. Doch bereits vorher sind einige wichtige Regeln zu beachten: Testautoren sind angehalten, die durch das Testformat (Publikation von Modellprüfungen) und Lernzielinventare gesetzten Limits genauestens einzuhalten. Ein zusätzliches Hilfsmittel sind darüber hinausgehende Richtlinien für Testkonstrukteure, die beispielsweise den Aufgabentyp und Textinput (z. B. hinsichtlich Länge, Textsorte etc.) genauer erläutern. Nach der Vorlage des Rohentwurfs durch den Autor kommt die Begutachtungs- und Redaktionsphase, in der der Test sich von außen weiteren kritischen Fragen stellen muss, die bereits dem Autor bekannt waren, die dieser aber möglicherweise übersehen hat: Ist ein Text für die Zielgruppe Migranten überhaupt geeignet? Gibt es Multiple-Choice-Items, die schlecht formuliert sind bzw. mehr als eine richtige Lösung ermöglichen? Gibt es Aufgaben, die auch ohne Lektüre des Textes lösbar sind? Gibt es Distraktoren, die so unplausibel sind, dass sie nach aller Wahrscheinlichkeit von niemandem gewählt werden? Nach Itemerstellung, Begutachtung und redaktioneller Überarbeitung erfolgt als weitere objektive Überprüfung der Güte des Testitems die Erprobung, die normalerweise mit einer Population von mindestens 100 Lernenden des Zielniveaus und der betreffenden Zielgruppe (in diesem Fall also Migranten) durchgeführt wird. Die statistische Analyse der Testitems (meist als klassische Itemanalyse mit Hilfe des Computerprogramms *Iteman* aus *MicroCAT*) führt dann für die Redaktion zu nützlichen Hinweisen bezüglich ihrer Qualität. Hier sei beispielsweise auf die Parameter „Schwierigkeit“ und „Trennschärfe“ verwiesen: Die Schwierigkeit eines Items wird ausgedrückt als Anteil der korrekten Antworten an der Gesamtzahl der Kandidatenantworten bei einem Item (*proportion correct*). Bei Erprobungen des *Zertifikat Deutsch* liegt die angemessene Schwierigkeit zwischen 0.50 und 0.90 prop. correct. D. h., dass ein Item in der Regel verworfen wird, wenn weniger als 50 oder mehr als 90 Prozent der Probanden es richtig lösen. Einzelne leichtere Aufgaben werden häufig aber auch als Einstiegsitems in einen Subtest akzeptiert. Bei der Untersuchung der Trennschärfe wird festgestellt, wie gut ein Item die schwachen von den starken Kandidaten trennt. Um die Trennschärfe zu beschreiben, werden der Trennschärfindex (*discrimination index*) und die punkt-biseriale Korrelation (*point biserial correlation*) errechnet. Weitere Aufschlüsse über die Qualität eines Items kann man daraus ziehen, ob die Distraktoren angemessen funktionieren. Wenn in einem Multiple-Choice-Item die Lösungsverteilung $a=0.19/b=0.64$ richtige Lösung/ $c=0.17$ ist, so funktioniert die Aufgabe mit zwei guten Distraktoren a und c recht gut. Ergibt die Analyse für eine andere Aufgabe die Werte $a=0.94$ richtige Lösung/ $b=0.04/c=0.02$, so wird die Testredaktion das Item überarbeiten müssen oder ganz verwerfen. Das Item ist erheblich zu leicht, die Distraktoren b und c erfüllen nicht ihren Zweck.

Es konnte an dieser Stelle nur ein kurzer Einblick in die fachwissenschaftliche Arbeit der Testentwicklung gegeben werden. Ganz allgemein sollen die exemplarisch aufgezeigten Prozeduren aber zeigen, dass gerade in dem sensiblen Bereich der Ü-

berprüfung von Sprachkompetenzen bei Migranten und Einbürgerungsbewerbern eine fachlich fundierte Herangehensweise angezeigt ist.

Zum Schluss sei noch ein kurzer Ausblick auf den *Deutsch-Test für Zuwanderer* gegeben, der ab 2009 das *Zertifikat Deutsch* als Abschlusstest in Integrationskursen ablösen soll. Die skalierte Prüfung (A2/B1) wird zurzeit gemeinsam von telc und dem Goethe-Institut entwickelt, das Testformat bis Ende 2007 bundesweit im Rahmen der oben beschriebenen Prozeduren erprobt. Als Prüfungsentwickler kann man sich dabei auf eine wichtige Vorarbeit stützen, die die Sprachbedarfe der Zielgruppe genauer erforscht hat.⁶ Das Ergebnis der Rechercharbeiten hat gezeigt, dass „Leben mit Kindern“ (96 Prozent), „Behörden“ (74 Prozent), „Arzt“ (56 Prozent), „Arbeits-suche“ und „Arbeitsplatz“ (jeweils 30 Prozent) als besonders wichtige Handlungsbe-reiche angesehen werden. „Briefliche Kommunikation“, „Medien“ oder „Feiern“ rangieren mit nur 1 oder 2 Prozent-Nennungen ganz hinten.

Für das Projektteam von telc und Goethe-Institut wird es wichtig sein, die Ergebnisse der Bedarfsanalyse so in ihre Entwicklungsarbeiten mit einzubeziehen, dass auch ab 2009 ein faires Testverfahren für die Feststellung der Sprachkompetenzen von Migranten verfügbar ist. Dabei ist der Anspruch, der durch die sprachliche Wirklichkeit mit der Lebenswelt der Migranten verbunden ist, genauso legitim wie die Ansprüche, die die Aufnahmegesellschaft mittels einer staatlichen Sprachkursförderung und einem damit verknüpften Prüfungskonzept an die Teilnehmenden stellt.

Anmerkungen

- 1 Weitere Informationen in www.telc.net – Siehe auch *Zertifikat Deutsch. Lernziele und Testformat*, hg. von WBT Weiterbildungs-Testsysteme [seit 2007 telc], Goethe-Institut, Österreichisches Sprachdiplom und Konferenz der Schweizerischen Erziehungsdirektoren, Frankfurt am Main, 1999.
- 2 Da die Papiere der Arbeitsgruppe zurzeit noch ALTE-intern diskutiert werden, kann hier leider nicht aus ihnen zitiert werden. Eine Veröffentlichung ist für die zweite Jahreshälfte 2008 geplant.
- 3 Die deutsche Übersetzung ist zugänglich als *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*, von John Trim, Brian North, Daniel Coste. München: Langenscheidt, 2001.
- 4 Ebenda, S. 81f.
- 5 Vgl. *Gemeinsamer europäischer Referenzrahmen*, S. 76.
- 6 Konrad Ehlich, Recherche und Dokumentation hinsichtlich der Sprachbedarfe von Teilnehmenden an Integrationskursen DaZ, München 2007.