COUNCIL OF EUROPE CONSEIL DE L'EUROPE

Language Policy Division
Division des Politiques linguistiques

Sprachtests für gesellschaftlichen Zusammenhalt und Staatsbürgerschaft – ein Leitfaden für Entscheidungsträger

Autoren

Andrew Balch
Michael Corrigan
Sara Gysen
Henk Kuijper
Michaela Perlmann-Balme
Shalini Roppe
Heinrich Rübeling
Steffi Steiner
Piet Van Avermaet
Beate Zeidler

Mit tatkräftiger Unterstützung der ALTE Arbeitsgruppe „Language Assessment for Migration and Integration“.

Dieser Leitfaden wurde im Rahmen eines Seminars in Straßburg 2008 vom Europarat als Teil der „Thematic Studies on the linguistic integration of adult migrants” veröffentlicht (www.coe.int/lang).

- "Thematic Studies"
- "Case Studies"
- "Seminar”

(1) Einleitung

Viele europäische Länder führen derzeit Anforderungen an die sprachlichen Kompetenzen für Menschen ein, die in das Land einwandern, sich niederlassen und die Staatsbürgerschaft erwerben möchten. Nationale Regierungen verlangen oftmals, dass für diesen Zweck Sprachtests oder andere formelle Verfahren der Leistungsmessung eingesetzt werden.

Ziel dieses Kapitels ist nicht, für den Einsatz solcher Sprachtests zu plädieren, sondern dort, wo Tests bereits verpflichtend sind oder eingeführt werden sollen, eine fachliche Beratung anzubieten, die sich auf bewährte Testverfahren stützt, um so sicher zu stellen, dass die Anforderungen der Testabnehmer berücksichtigt werden und die Tests für die Teilnehmenden fair sind. Testfairness ist eine besonders wichtige Qualität, wenn es sich um Tests handelt, die zum Zwecke der Einwanderung, der Niederlassung oder der Staatsbürgerschaft eingesetzt werden. Unfaire Tests können zur Folge haben, dass Migrantinnen und Migranten Bürger- und Menschenrechte verweigert werden. Es gibt eine Reihe von leicht zugänglichen Standards (siehe *weiterführende Lektüre*), die Hilfen bei der Entwicklung und Durchführung von fairen Tests anbieten. Sie können die Lektüre dieses Kapitels ergänzen und beispielhaft aufzeigen, wie sich die verschiedenen Phasen des Testerstellungsprozesses von den ethischen Grundlagen ableiten.

Entscheidungsträger, die die Einführung eines Sprachtests erwägen, sollten in jedem Fall zunächst über die folgenden Fragestellungen nachdenken:

- Ist es sinnvoller, eine andere Form der Leistungsmessung als einen Sprachtest zu wählen?
- Könnte es sinnvoll sein, verschiedene Formen der Leistungsmessung zu kombinieren?
- Welcher Gebrauch wird von den Testergebnissen gemacht?
- Was werden die Konsequenzen des Tests für die Gesellschaft sein?
- Was werden die Auswirkungen für den Migranten/die Migrantin sein?
- Was werden die Auswirkungen für die Gesellschaft der Migrantin/des Migranten sein?

Art der Leistungsmessung

Bei einer Erörterung der ersten und zweiten Frage oben sollten Entscheidungsträger wissen, dass es neben Tests andere Formen der Leistungsmessung gibt, die auch angemessen sein könnten. Jede Form hat ihre eigenen Vorteile hinsichtlich gewisser Merkmale wie Auswirkungen auf den Kandidaten, Rückschlüsse aus den Testergebnissen, Standardisierung und Reliabilität der Ergebnisse, Kosten und Praktikabilität. Es ist deshalb wichtig, über die Anforderungen der Situation und der Gegebenheiten nachzudenken, um die am besten geeignete Form der Leistungsmessung zu finden. Und es sollte nicht vergessen werden, dass verschiedene Formen miteinander kombiniert werden können. Im Folgen sollen einige Vorteile von Tests und anderen Formen der Leistungsmessung genannt werden.

Tests, die angemessen entworfen, erstellt und durchgeführt werden, haben folgende Vorteile:

- Die Ergebnisse sind weitestgehend standardisiert und reliabel. Das bedeutet, dass es einfach ist Kandidaten miteinander zu vergleichen, die die Prüfung zum gleichen oder einem späteren Zeitpunkt abgelegt haben.
- Die Leistung des Kandidaten wird mit einem hohen Grad an Objektivität gemessen.
- Man kann in kurzer Zeit eine große Anzahl an Kandidaten prüfen.

Eine Alternative zu Tests könnte darin bestehen, während des Kurses laufend Beurteilungen vorzunehmen, oder vom Kandidaten eingereichte Arbeiten zu bewerten.[1] Sollte die Leistungsmessung eine stark formative Funktion haben, kann sie in den Unterricht integriert werden und so dem Lernenden helfen, mehr Verantwortung für seinen Lernprozess zu übernehmen. Die Beurteilungen können sowohl durch einen Mitschüler als

auch durch die Gruppe erfolgen. Zusätzlich zur formativen Funktion solcher Beurteilungen, gibt es weitere Vorteile dieses Ansatzes:

- Die Leistungen für die Beurteilung können in einer nicht-bedrohlichen Atmosphäre (z.B. dem Klassenzimmer) erhoben werden, was ihre Validität als Nachweis der wirklichen Leistung des Kandidaten oder der Kandidatin erhöhen könnte.
- Leistungen können durch lebensnahe Aufgaben erhoben werden oder den Aufgaben entsprechen, die Migrantinnen und Migranten im realen Leben zu bewältigen haben.

[1] Vgl. das Kapitel von David Little zum Sprachenportfolio in diesem Band

- Er bietet eine bessere Möglichkeit, die Leistung des Kandidaten oder der Kandidatin holistisch zu bewerten und damit deren zugrunde liegende Fähigkeit zu erfassen, Aufgaben erfolgreich zu lösen, da weniger Gewicht auf das Überprüfen einzelner sprachlicher Elemente gelegt wird.

Auswirkungen
Was die Schlussfolgerungen betrifft, die aus den Testergebnissen gezogen werden, sollte man bedenken, dass bei jedem Test die Ergebnisse immer mit einem gewissen Spielraum für mögliche Fehler errechnet werden, weil keine Leistungsmessung von sich behaupten kann, ohne Messfehler zu sein. Dies Kapitel dient deshalb dazu, die Möglichkeiten negativer Schlussfolgerungen aus dem Einsatz eines Tests zu *minimieren,* indem es unter anderem die Fehlermöglichkeiten reduziert, statt zu versuchen sie ganz auszuschalten. Darüber hinaus können Tests mit hohen Kandidatenzahlen nur schwer die individuelle Persönlichkeit, die Lernerbiografie und die persönliche Geschichte eines jeden Lernenden berücksichtigen. Wenn ein Kandidat in einem Test schlechte Leistungen erbracht hat, kann seine wirkliche Fähigkeit unterschätzt werden. Der Gesamtnutzen, den man sich vom Einsatz eines bestimmten Tests erhofft, muss daher in Relation zu den Konsequenzen bei Nichtbestehen des Tests gesehen werden, da manche auf den Testergebnissen basierende Entscheidungen falsch sein können.

Wenn über die Art der Leistungsmessung entschieden wurde, ist es sehr wichtig, über den Gebrauch und die Konsequenzen nachzudenken, da der Gebrauch weit reichende und unerwartete Konsequenzen nach sich ziehen kann. Es ist deshalb ratsam, sich bei der Planung der Leistungsmessung die möglichen Konsequenzen genau zu überlegen, und bei der Durchführung der Leistungsmessung zu untersuchen, welches die tatsächlichen Konsequenzen sind. Mögliche Konsequenzen des Tests können eine Veränderung der Lehr- und Lernverfahren sein oder Veränderungen im Erziehungssystem des Heimatlandes der Migrantinnen und Migranten.

Wenn ein Sprachtest eingesetzt werden soll, ist es aufgrund der oben genannten Überlegungen notwendig, dass alle Beteiligten – Entscheidungsträger eingeschlossen - sich sicher sind, dass (i) der Test entwickelt wurde, um den festgestellten Bedürfnissen zu entsprechen, und dass (ii) er so funktioniert wie beabsichtigt, damit die damit verbundenen Ziele angemessen und fair umgesetzt werden können. Testfairness ist maßgeblich für alle Arten von Tests und für alle Zielgruppen. Sie erhält jedoch im Kontext der Sprachprüfungen für Migration, Niederlassung und Aufenthalt sowie Staatsbürgerschaft eine besondere Bedeutung aufgrund der erheblichen Konsequenzen für die Teilnehmenden im Hinblick auf Bürger- und Menschenrechte. Die Maßnahmen, die sicherstellen, dass ein Test fair ist, beginnen in der Planungsphase des Tests und setzen sich in allen weiteren Phasen der Erstellung und Durchführung fort. Diese Maßnahmen erlauben den Testbenutzern, die Ergebnisse eines Tests angemessen zu interpretieren und zu verwenden. Um diesbezüglich die Entscheidungsträger in ihrer Verantwortlichkeit zu unterstützen, soll dieses Kapitel Informationen über alle Phasen der Testentwicklung und Durchführung von Tests bereitstellen und Argumente liefern hinsichtlich der Auswahl und Überwachung des Testanbieters sowie der Interpretation der Ergebnisse. Außerdem könnte es als Leitfaden dienen bei der Entwicklung und Erstellung eigener Tests. Die Anwendung der in diesem Kapitel beschriebenen bewährten Testverfahren stellt bei allen Tests - auch bei Tests im Kontext der Migration - sicher, dass nicht nur geeignete Fähigkeiten und geeignetes Wissen geprüft werden (Testvalidität), sondern dass dies auch für alle Kandidaten gleichermaßen und in allen Prüfungssätzen der Fall ist (Testreliabilität). Außerdem verweist das Kapitel auf weitere Quellen, die sich als hilfreich erweisen könnten.

Weiterführende Literatur:

ALTE Code of Practice – http://www.alte.org/quality_assurance/index.php

Multilingual Glossary of Language Testing Terms, Studies in Language Testing volume 6, Cambridge University Press (ISBN: 0-521-65877-2)

ILTA Code of Ethics – http://www.iltaonline.com/code.pdf

JCTP Code of Fair Testing Practice in Education – http://www.apa.org/science/FinalCode.pdf

(2) Entscheiden, was getestet werden soll

(2.1) Übersicht

In diesem Abschnitt werden die Schritte erläutert, die dazu dienen, die Angemessenheit des Tests im Hinblick auf seinen Verwendungszweck sicherzustellen. Der erste Schritt in diesem Prozess ist die präzise und eindeutige Ermittlung von Zweck und Ziel des Tests. Anschließend folgt die Festlegung der Inhalte und der Schwierigkeit. Zum Schluss muss eine Testbeschreibung (engl. *test specification*) entwickelt werden, ein unerlässliches Dokument für die späteren Phasen der Testentwicklung und -revision.

(2.2) Festlegung des Testziels und der Anforderungen an die Teilnehmenden

Vor der Entwicklung jeglicher Sprachtests muss als erstes der genaue Verwendungszweck bestimmt werden. Die Angabe „Test für Migration und Staatsbürgerschaft“ reicht allein nicht aus, denn in diesem Bereich gibt es eine weite Bandbreite von Gründen für das Testen von Migranten. Diese reichen von Lernermotivation (den Lernenden helfen, ihre gegenwärtige Sprachkompetenz in der Zielsprache zu nutzen und zu verbessern) und dem Erfassen der Sprachkompetenz im Hinblick auf die Partizipation in klar definierten sozialen Situationen (z.B. Studium oder Arbeit) bis zu Entscheidungen, die die Rechtsansprüche dieser Zielgruppe beeinflussen, wie zum Beispiel ihr Recht in einem Land zu bleiben oder dessen Staatsbürgerschaft anzunehmen.

Erst wenn der Zweck klar definiert wurde, ist es möglich, die Anforderungen zu identifizieren, denen der Teilnehmende in der Realsituation begegnen wird und die sich im Test widerspiegeln sollten wie z.B. die Notwendigkeit, sich an gesellschaftlichen Prozessen zu beteiligen und die staatsbürgerlichen Rechte und Pflichten auszuüben. Ein klar definierter und transparenter Verwendungszweck trägt nicht nur zur Testfairness bei, indem er die Anforderungen verdeutlicht, sondern ermöglicht den Beteiligten auch, die Testergebnisse angemessen zu interpretieren und zu nutzen. Dieser Prozess der Festlegung der Anforderungen in einer Realsituation wird als Bedarfsanalyse bezeichnet.

Diese Bedarfsanalyse sollte auch die Tatsache berücksichtigen, dass es verschiedene Untergruppen von Zuwanderern mit eigenen, speziellen Bedürfnissen gibt. Diejenigen zum Beispiel, die sich so schnell wie möglich in den Arbeitsmarkt integrieren möchten, haben oftmals andere Bedürfnisse als diejenigen, die zu Hause Kinder großziehen. Für eine Bedarfsanalyse hat sich die Verfahrensweise von Sprachtestentwicklern bewährt, erst einmal die für die Zielgruppe relevanten Kontexte und Situationen genau zu bestimmen. Bei der Planung solcher Bedarfsanalysen sollten Entscheidungsträger sicherstellen, dass dafür ausreichende Mittel zur Verfügung stehen und dass Vertreter verschiedener Gesellschaftsgruppen bei der genauen Beschreibung der Bedürfnisse einbezogen werden.

(2.3) Festlegung der sprachlichen Anforderungen

Wenn die Anforderungen in der Realsituation identifiziert wurden, müssen sie in sprachliche Anforderungen umgesetzt werden, die nicht nur die Kenntnisse und die Fertigkeiten präzisieren, sondern auch das dafür notwendige Kompetenzniveau. Wenn zum Beispiel ein Sprachtest dafür entwickelt wurde zu messen, ob der Teilnehmende über die nötige Sprachkompetenz verfügt, um an einem Kurs zur Berufsausbildung teilzunehmen, könnte man erwarten, dass der Teilnehmende über die Fähigkeit verfügt, Unterricht und Seminaren zu folgen, mit Lehrkräften und Mitschülern zu kommunizieren, relevante Fachliteratur zu lesen und schriftliche Aufgaben zu erledigen.

Diese Analyse könnte dann helfen, das erforderliche Sprachniveau zu bestimmen beziehungsweise das erforderliche Sprachniveau in jeder der vier getesteten Fertigkeiten, zum Beispiel im Lesen und Schreiben. Wenn dagegen ein Sprachtest zur Zertifizierung von Übersetzern oder Dolmetschern entwickelt wurde, dann würde eine vergleichbare

Bedarfsanalyse des Berufsbilds zeigen, dass das erforderliche Sprachniveau sehr viel höher wäre. Zudem würde man herausfinden, dass im Falle des Übersetzers ein höheres Niveau in den Fertigkeiten Lesen und Schreiben verlangt würde, während für den Beruf des Dolmetschers die mündliche Fähigkeit von größerer Bedeutung wäre.

Im Gegensatz zu den oben beschriebenen Beispielen lassen sich die sprachlichen Anforderungen von Zuwanderern und Bewerbern um die Staatsbürgerschaft nicht so eindeutig aus der Realsituation ableiten. Die Verbindung zwischen dem Sprachniveau in der offiziellen Sprache bzw. den offiziellen Sprachen und der Fähigkeit zur Integration in die Gesellschaft und/oder zur Ausübung der mit der Staatsbürgerschaft einhergehenden Rechte und Pflichten ist sehr viel schwieriger zu bestimmen. Wenn das Sprachniveau wirklich der einzige Einflussfaktor wäre, dann wären alle Staatsbürger eines Landes voll integriert. Da dies nicht der Fall ist, liegt die Schlussfolgerung nahe, dass auch andere Faktoren wichtig sind. Die Aufgabe des Testentwicklers besteht dennoch darin, die erforderlichen sprachlichen Anforderungen zu identifizieren. Nach der genauen Bedarfsanalyse ist es auch wichtig sicherzustellen, dass keine falschen Annahmen über die kulturelle Herkunft der Kandidaten und Kandidatinnen oder deren Bildungshintergrund den Test beeinflussen. Sprachprüfungen für Studien- oder Arbeitszwecke werden in den meisten Fällen von Kandidatengruppen abgelegt, die im Hinblick auf ihren Bildungshintergrund und ihre kognitiven Fähigkeiten homogen sind. Tests für Integration und Staatsbürgerschaft, also Tests zur Erlangung von Bürgerrechten, müssen dagegen darauf ausgerichtet sein, eine große Bandbreite von möglichen Kandidaten abzudecken, d.h. sie müssen sowohl für Menschen mit geringem Bildungsgrad als auch für solche mit einem hohen akademischen Bildungsgrad zugänglich sein.

(2.4) Bestimmung des angemessenen Schwierigkeitsgrades

Nachdem die sprachlichen Anforderungen identifiziert wurden, müssen die Testentwickler versuchen, diese den Kannbeschreibungen des Gemeinsamen europäischen Referenzrahmens für Sprachen (GER) zuzuordnen, der erstmals 2001 veröffentlicht wurde:

Dieser Referenzrahmen

> *"stellt eine gemeinsame Basis dar für die Entwicklung von zielsprachlichen Lehrplänen, curricularen Richtlinien, Prüfungen, Lehrwerken usw. in ganz Europa. Er beschreibt umfassend, was Lernende zu tun lernen müssen, um eine Sprache für kommunikative Zwecke zu benutzen, und welche Kenntnisse und Fertigkeiten sie entwickeln müssen, um in der Lage zu sein, kommunikativ erfolgreich zu handeln."*[2]

Der Referenzrahmen enthält eine Reihe von illustrativen Skalen (jeweils für Sprechen, Schreiben, Lesen, Hören sowie für Interaktion), die Kompetenzniveaus

definieren und zeigen, was ein Lernender auf jedem Niveau sprachlich realisieren kann. Sie erlauben es, den Lernfortschritt auf einer sechsstufigen Niveauskala zu

---

[2] Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen, im Auftrag des Europarats, Rat für kulturelle Zusammenarbeit, deutsche Ausgabe Hg. vom Goethe-Institut Inter Nationes, der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK), der Schweizerischen Konferenz der Kantonalen Erziehungsdirektoren (EDK) und dem österreichischen Bundesministerium für Bildung, Wissenschaft und Kultur (BMBWK), München, Langenscheidt 2001, S.14.

messen, die von A1 (elementare Sprachverwendung) bis zu C2 (kompetente Sprachverwendung) reicht.

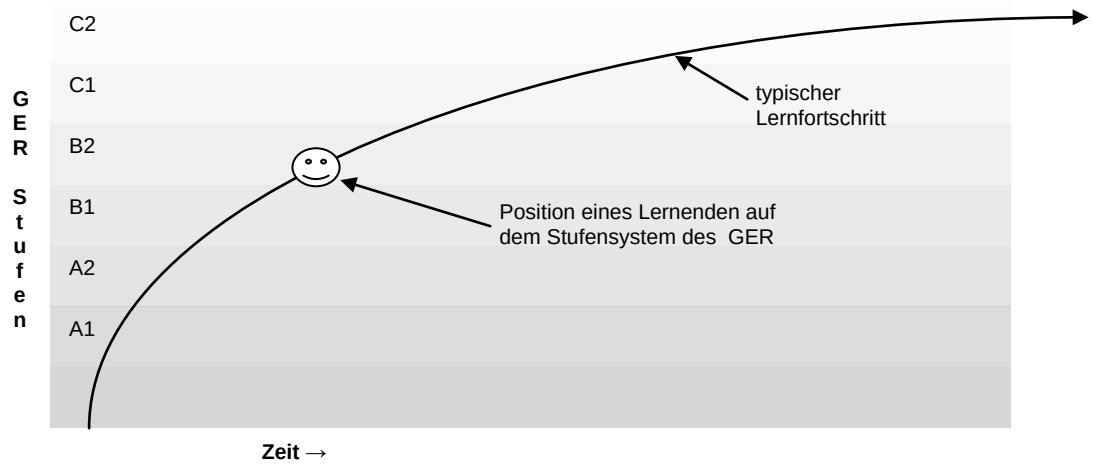

Tabelle 1 Lernfortschritt und GER Kompetenzniveaus

Tabelle 1 zeigt eine typische Lernfortschrittskurve auf der horizontalen Achse entlang der Stufen des GER. Dabei ist zu beachten, dass der gleiche Abstand von einer Stufe zur nächsten nicht bedeuten soll, dass eine ähnliche Zeitspanne benötigt wird, um die nächsthöhere Niveaustufe zu erreichen. Deshalb ist die Lernfortschrittskurve am Anfang steil und flacht gegen Ende ab. Das liegt daran, dass die Bandbreite von Fähigkeiten und Sprachmitteln von Stufe zu Stufe zunimmt und damit auch die benötigte Zeit für den Fortschritt von einer Niveaustufe zur nächsten. Eine Bedarfsanalyse sollte somit nicht auf der Stundenzahl basieren, die für den Lernerfolg notwendig ist, sondern auf dem sorgfältigen Abgleich zwischen den Anforderungen der Testabnehmer und den Kannbeschreibungen des Referenzrahmens.

Testentwickler sollten sich zudem darüber im Klaren sein, dass die Fähigkeiten der meisten Lernenden nicht so gleichmäßig über die vier Fertigkeiten Lesen, Hören, Sprechen und Schreiben verteilt sind, wie Tabelle 1 das suggerieren könnte. Vielmehr sind die Kompetenzen in der mündlichen Kommunikation oft besser entwickelt als im Schreiben und Lesen. Der Lernende kann deshalb im Hinblick auf die vier Fertigkeiten ein *gezacktes* Profil haben, wie Tabelle 2 zeigt.

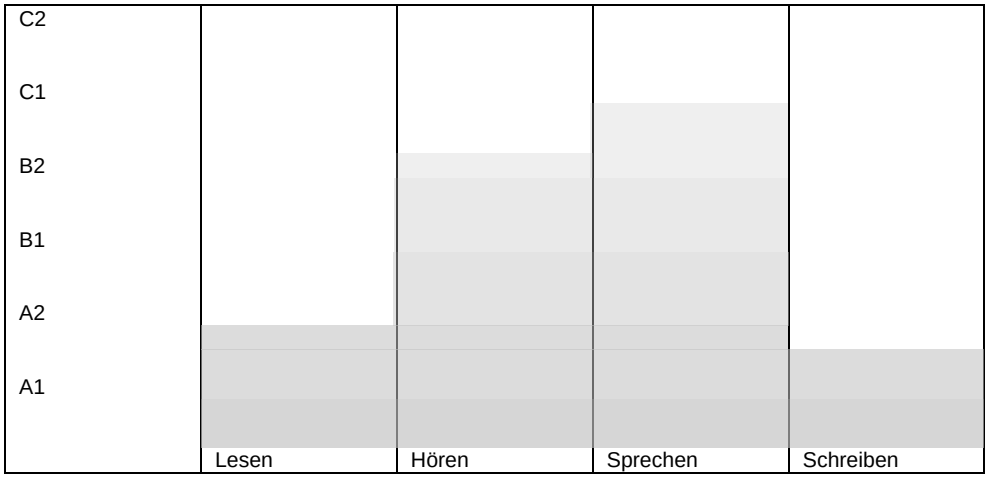

Tabelle 2 Gezacktes Profil eines Lernenden

Ein modularer Ansatz beim Testen hat - neben einer genaueren Abbildung der Fähigkeiten eines Kandidaten oder einer Kandidatin – mehrere Vorteile. Wenn es möglich ist, die

Testteile zu den vier Fertigkeiten einzeln zu absolvieren, können die Testteilnehmer und Teilnehmerinnen statt aller vier Testteile zunächst nur diejenigen Testteile absolvieren, die sie besser beherrschen, was zu einer höheren Motivation führt. Diese Möglichkeit wäre besonders für Migrantinnen und Migranten geeignet, die nicht schreiben oder lesen können, oder für Teilnehmende mit sehr wenig Schreiberfahrung. Ein positiver Nebeneffekt dieses Ansatzes besteht darin, dass der Sprachunterricht sich auf die Fertigkeit konzentrieren könnte, die besonderer Aufmerksamkeit bedürfen. Wenn die Ergebnisse aus solchen Tests zu Teilkompetenzen von den Testinstitutionen kommuniziert werden, muss aber sorgfältig darauf geachtet werden, dass die Ergebnisse nicht verwechselt werden mit denen aus einem Test, der Aufgaben zu allen vier Fertigkeiten enthält. Zeugnisse sollten daher nicht nur eine Gesamtniveaustufe angeben, sondern vielmehr die Niveaustufe, die in jeder einzelnen Fertigkeit erreicht wurde.

Weiterführende Literatur:

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment – http://www.coe.int/t/dg4/linguistic/CADRE_EN.asp

(2.5) Anfertigung von Testbeschreibungen

Sobald die Zielgruppe, der Verwendungszweck und die Zielsetzung des Tests bestimmt worden sind, sollten diese in einer detaillierten Testbeschreibung (*test specification*) schriftlich festgehalten werden. Die Testbeschreibung enthält auch den Item- oder Aufgabentyp, der benutzt wird, das Testformat und Hinweise zur Durchführung des Tests. Die Hinweise zur Durchführung werden von den Zielen des Tests und der Zielgruppe abgeleitet. Diese Beschreibung fungiert dann als Referenzdokument, das Informationen für alle Entscheidungen in einer späteren Phase des Testeinsatzes bereithält.

(3) Sicherstellen, dass die Testbeschreibung in der Praxis umgesetzt wird

(3.1) Überblick

Nach der Fertigstellung der Testbeschreibung sind weitere Schritte nötig, damit der Test wie beabsichtigt funktioniert. Bewertungskriterien und das Testformat müssen entwickelt werden, Testaufgaben müssen gemäß den Spezifikationen geschrieben werden und diese Aufgaben müssen dann in geeigneter Weise, wie in der Testbeschreibung dargelegt, zu einem ganzen Test kombiniert werden. Der so produzierte Test muss zuverlässig und fair für die potenziellen Teilnehmenden durchgeführt werden. Schließlich müssen die Daten der Testdurchführung analysiert werden, um zu bestätigen, dass der Test wie erwartet funktioniert. Sollte dies nicht der Fall sein, müssen an den entsprechenden Stellen Anpassungen vorgenommen werden. Während dieses ganzen Prozesses sind Qualitätskontrollen notwendig. Die am Ende der Einleitung genannten Veröffentlichungen helfen Testanbietern, die professionellen und ethischen Standards in der Praxis umzusetzen. Die weiterführende Literatur am Ende dieses Kapitels beschäftigt sich ausführlicher mit diesen Standards.

Weiterführende Literatur:

AERA/APA/NCME Standards for Educational and Psychological Testing – http://www.apa.org/science/standards.html

ALTE COP QMS Checklists – http://www.alte.org/quality_assurance/code/checklist.php

ALTE Minimum standards for establishing quality profiles in ALTE examinations – http://www.alte.org/quality_assurance/index.php

ALTE Principles of Good Practice – http://www.alte.org/quality_assurance/code/good_practice.pdf

EALTA Guidelines for Good Practice – http://www.ealta.eu.org/guidelines.htm

ILTA Draft Code of Practice –

http://www.iltaonline.com/ILTA-COP-ver3-21Jun2006.pdf

(3.2) Bewertungskriterien und Testformat

Die in den Testbeschreibungen (Kapitel 2.5) skizzierte Zielsetzung des Tests muss in spezifische, einzeln zu testende Aspekte untergliedert werden, damit der Testentwickler einen Nutzen davon hat. Erst dann, wenn präzisiert wurde, wie die Kandidatenleistung bewertet wird, kann die Entwicklung einer angemessenen Kombination von Aufgaben und Aufgabentypen beginnen. Der Testentwickler sollte vor allem berücksichtigen, dass Kandidaten angemessene Aufgaben benötigen, damit sie zeigen können, dass sie die Leistungsanforderungen erfüllen.

(3.3) Itemerstellung

Um geeignete Testitems zu erstellen, müssen die Testautoren klare Richtlinien bekommen. Diese sollten normalerweise einen Überblick über die Zielgruppe und das Testziel bieten sowie allgemeine Hinweise enthalten zu geeigneten und ungeeigneten Themen, zum Input (zum Beispiel die Anzahl der Wörter für Lesetexte) und zum Output (zum Beispiel die Anzahl der Wörter, die ein Kandidat schreiben sollte), zum Grad der „Authentizität“ der Texte, um nur einige Beispiele zu nennen. Erst auf diesem Hintergrund kann jedes einzelne Item genauer geprüft werden. Wenn die Items fertig gestellt worden sind, sollten Experten beurteilen, ob die Testaufgaben die Vorgaben der Testbeschreibung angemessen umsetzen.

Weiterführende Lektüre:

Item Writer Guidelines – http://www.alte.org/projects/item-writer.php (dort die Übersetzung ins Deutsche)

3.4 Erprobungen

Nach der Expertenbegutachtung (vgl. Kapitel 3.3) sollten Items und Aufgaben vorliegen, die für die Prüfung geeignet erscheinen. Dennoch ist eine Bestätigung notwendig, dass die Items tatsächlich so funktionieren wie beabsichtigt - dass sie also die Zielsprachenkompetenz prüfen, stärkere von schwächeren Teilnehmenden wirksam unterscheiden können, einzelne Kandidatengruppen nicht benachteiligen usw. Dafür muss das Material unter echten Testbedingungen und mit Teilnehmenden erprobt werden, die hinsichtlich ihres demographischen Profils der tatsächlichen Testpopulation so ähnlich wie möglich sind. Die Ergebnisse einer Erprobung von geschlossenen Aufgaben (z.B. Multiple-Choice) können einer detaillierten statistischen Analyse unterworfen werden, die Analyse von offenen Aufgaben (z.B. Aufgaben zum Sprechen) hingegen geschieht qualitativ durch Expertenurteil, um zu sehen, in welchem Maße die Performanz der Teilnehmenden den Erwartungen der Testersteller entspricht.

Aufgrund solcher Analysen werden die Items und Aufgaben entweder für echte Prüfungen verwendet, oder überarbeitet und nochmals erprobt oder endgültig verworfen. Wenn darüber hinaus eine *Itembank* besteht, in der die itembezogenen Erprobungsergebnisse gespeichert werden, so können aus dieser Itembank Items und Aufgaben für zukünftige Prüfungen zusammengestellt werden, die bestimmten vordefinierten Anforderungen entsprechen, z.B. eine bestimmte Schwierigkeitsstufe abbilden.

Es ist wichtig, dass Experten die Texte und die Items an verschiedenen Stellen des Erstellungs- und Erprobungsprozesses überprüfen. Außerdem sollte diese Überprüfung mehr als einmal stattfinden, da Aufgaben, wenn sie geändert werden, noch einmal begutachtet werden müssen, und nach einer der Erprobung neue Informationen zur Verfügung stehen. Einige Instrumente, die bei dieser Überprüfung hilfreich sind, werden unten in der weiterführenden Literatur aufgelistet. Mit Hilfe der dort genannten Kategorien können Aufgaben, Aufgabengruppen und in einigen Fällen auch die Teilnehmer-Lösungen analysiert werden. Auf diese Weise können Unterschiede zwischen einzelnen Aufgaben bzw. zwischen Aufgabe und Testbeschreibung sichtbar werden. Die u.a. Raster wurden ursprünglich erstellt, um Prüfungsanbietern die Positionierung ihrer Prüfungen auf den Niveaustufen des GER zu erleichtern, können jedoch auch für den beschriebenen Zweck benutzt werden.

Weiterführende Literatur:

Content Analysis Checklists (Checklisten zur Analyse von Prüfungsinhalten) – http://www.alte.org/projects/content.php

CEFR Grids for the analysis of test tasks (listening, reading, speaking and writing) - http://www.coe.int/T/DG4/Portfolio/?L=E&M=/documents_intro/Manual.html

3.5 Testdurchführung

Testanbieter müssen sicherstellen, dass die Prüfung unter Bedingungen durchgeführt wird, die für alle Teilnehmenden gleichermaßen fair sind. Aus diesem Grund wird empfohlen, den Testablauf genau zu regeln, damit die Unterschiede in der Durchführung des Tests so gering wie möglich sind. Die Bestimmungen zur Durchführung sollten Folgendes enthalten:

- Alle Prüfungszentren werden für die Testdurchführung ordnungsgemäß akkreditiert.
- Die Mitarbeiter im Prüfungszentrum sind ausgebildet und gut informiert, erhalten die notwendige Unterstützung, und die Testdurchführung wird ausreichend kontrolliert.
- Die Prüfungszentren gewährleisten während des gesamten Prozesses von der Prüfungsanmeldung bis zur Mitteilung der Ergebnisse ein hohes Maß an Sicherheit und Geheimhaltung.
- Die Bedingungen, unter denen die Prüfung abgenommen wird, sind angemessen (Geräuschbelastung im Prüfungsraum, Temperatur, Abstand zwischen den Teilnehmenden usw.).
- Die Regelungen für Testteilnehmende mit besonderen Bedürfnissen.

(3.6) Berücksichtigung von Behinderten, Kranken und anderen Teilnehmenden mit besonderen Bedürfnissen

Das Testsystem darf Prüfungsteilnehmende mit besonderen Bedürfnissen nicht benachteiligen. Besondere Bedürfnisse können z.B. bedingt sein durch vorübergehende oder längerfristige körperliche, geistige oder emotionale Beeinträchtigungen, Lernbehinderungen und Lernstörungen sowie vorübergehende oder längerfristige Krankheiten. Auch Analphabetismus in der Mutter- oder Zielsprache, religiöse Vorschriften, eine Haftstrafe oder andere Rahmenbedingungen, können es für einen Prüfungsteilnehmenden schwierig oder unmöglich machen, in derselben Weise am Test teilzunehmen wie andere.

Es sollten Regelungen bestehen hinsichtlich

- der Freistellung von der Prüfung oder von einzelnen Prüfungsteilen,
- der Maßnahmen zur Sicherstellung einer fairen Beurteilung von Leistungen der o.g. Teilnehmenden,
- der Institution, die für die Entscheidung über eine Freistellungen von der gesamten Prüfung oder von einzelnen Prüfungsteilen zuständig ist,
- der speziellen Bedingungen im Einzelfall (z.B. Testunterlagen in Braille, Testunterlagen in Großdruck, Bereitstellung einer Braille-Schreibmaschine oder eines Computers mit Sonderausstattung, eines Vorlesers, Schreibers oder Assistenten, mehr Zeit für bestimmte Testteile, zusätzliche Pausen, Gebärdensprachendolmetscher, gesonderte Prüfungstermine oder Prüfungsorte),
- der Möglichkeit, Einspruch gegen die Entscheidungen einzulegen und wie über den Einspruch entschieden wird.

Alle diesbezüglichen Informationen sollten öffentlich verfügbar und für alle Prüfungsteilnehmenden zugänglich sein.

Weiterführende Literatur:

"Special Educational Needs in Europe. The Teaching & Learning of Languages. Teaching Languages to Learners with Special Needs", European Commission, DG EAC 23 03 LOT 3, January 2005.

(3.7) Bewerten und Benoten

Geschlossene Items können zuverlässig maschinell oder nach Anleitung ausgewertet werden. Offene Items hingegen müssen normalerweise von geschulten Prüfenden bewertet werden. In diesem Prozess können sowohl Übertragungsfehler als auch Bewertungsfehler auftreten. Bewertungsfehler äußern sich durch Inkonsistenzen in der Interpretation der Teilnehmerleistung und/oder der Bewertungskriterien des Tests. Inkonsistenz kann sich sowohl im Vergleich zu anderen Prüfenden als auch im Vergleich zur eigenen Bewertung über einen gewissen Zeitraum hinweg manifestieren. Sie kann durch gründliches Training sowie durch Überprüfung der Bewertungen stark verringert werden. Auch die Bewertung von Teilnehmerleistungen durch mehrere Bewerter kann diesen Effekt mindern. Bewertungen können zudem gewichtet werden, um konsistent auftretende besondere Milde oder Strenge auszugleichen. Im Falle starker Abweichungen müssen die Bewerter nachgeschult werden.

(3.8) Überwachung der Funktionalität des Tests

So wie man das Bewerterverhalten überwachen sollte (siehe Kapitel 3.7), ist es für Testentwickler notwendig, sowohl die Lösungen der Teilnehmenden als auch deren demografische Daten (z.B. Alter, Geschlecht und Nationalität) zu erheben und zu analysieren.

Damit wird sichergestellt, dass

- jeder Test die Fertigkeiten misst, die gemessen werden sollen,
- diese Fertigkeiten in allen Testversionen auf die gleiche Art und Weise gemessen werden,
- alle Testversionen für alle Teilnehmenden unabhängig von ihrer Herkunft faire Bedingungen bieten.

Die gewonnenen Erkenntnisse sollten dazu verwendet werden sicherzustellen, dass das Testergebnis die Fähigkeiten des Kandidaten zuverlässig abbildet. Außerdem sollten Erkenntnisse aus der Analyse an die Testkonstruktion, Testdurchführung und Bewertung rückgekoppelt werden, damit diese Prozesse kontinuierlich verbessert werden können.

(3.8.1) Analyse der Teilnehmerlösungen

Teilnehmerlösungen aus Echtprüfungen werden genutzt, um die Rohwerte des Tests zu errechnen und dem Testanbieter Informationen darüber zu liefern, wie gut ein Item oder eine Aufgabe die Kompetenz der Teilnehmenden misst (siehe Kapitel 3.4). Nicht nur die Schwierigkeit der Items, sondern auch ihre Trennschärfe (das Ausmaß, in dem ein Item dazu beiträgt, fähige von weniger fähigen Teilnehmenden zu unterscheiden, was ja das eigentliche Ziel des Tests ist) muss beachtet werden. Diese und alle anderen zum Test erhobenen Statistiken müssen aufbewahrt werden, um Vergleiche zwischen den Testversionen anstellen zu können. Dies soll sicherstellen, dass die Ergebnisse von verschiedenen Testversionen untereinander vergleichbar sind. Wenn eine Testentwicklung auf Erprobungen basiert, dann ist anzunehmen, dass Items und Aufgaben in der Echtprüfung wie erwartet funktionieren. Dennoch ist es wichtig, dies durch eine Analyse abzusichern. Wenn die Items nicht wie erwartet funktionieren, sollten die Ursachen ermittelt und die Version entsprechend überarbeitet werden.

Es ist selbstverständlich auch wichtig, dass jede Testversion die Kompetenz eines Kandidaten oder einer Kandidatin in derselben Weise misst wie vorherige und folgende Versionen. Aus diesem Grund werden häufig auch die Ergebnisse aus Echttests verwendet, um die Noten oder die Bestehensgrenze festzulegen. Wenn ein Kandidat zwei verschiedene Versionen ein- und desselben Tests absolvieren würde, wäre es unwahrscheinlich, dass er eine identische Punktzahl erreichen würde. Dennoch ist es möglich, mit einiger Gewissheit sicherzustellen, dass der Kandidat dieselbe Note erreichen würde oder bei beiden Versionen auf derselben Seite der Bestehensgrenze verortet würde. Dies wird durch eine *Angleichung* der Versionen ermöglicht, d.h. es wird eine Aussage darüber getroffen, welche Punktzahl in Testversion B gleichzusetzen ist mit der Punktzahl, die zum Bestehen von Version A benötigt wird. Wurde eine Erprobung durchgeführt, so ist dies leichter und präziser zu ermitteln.

(3.8.2) Überwachung von Bias

Ähnliche Vorgehensweisen wie oben in 3.8.1 beschrieben sollten auch auf die Leistungen von Kandidatengruppen auf gleichem Kompetenzniveau angewandt werden. Wenn zum Beispiel Kandidaten auf dem gleichen Niveau und mit derselben Herkunft im Vergleich zu anderen Teilnehmenden auf diesem Niveau im Test signifikant besser oder schlechter bei einem Item abschneiden, so könnte es daran liegen, dass dieses Item aus nichtsprachlichen Gründen zu Gunsten oder zum Nachteil von bestimmten Herkunftsgruppen funktioniert und deshalb einige Teilnehmende in unfairer Weise benachteiligt. Der Grund könnte jedoch auch in rein sprachbedingten Unterschieden oder Ähnlichkeiten zwischen Muttersprache und Zielsprache liegen, was dann nicht als unfair erachtet werden kann. Bei Verzerrungen ist deshalb eine qualitative Untersuchung notwendig. Wenn sich herausstellt, dass ein Item das Ergebnis aus nichtsprachlichen Gründen verzerrt, dann sollte es eliminiert werden. Gegebenenfalls sollte auch der Prozess der Itemerstellung überprüft werden, um solche Fälle für die Zukunft auszuschließen.

(4) Abschließende Bemerkungen

Der Gebrauch von Tests im Kontext von Integration und Einbürgerung ist erheblicher komplexer, als man zunächst vermuten würde. In diesem Artikel wurde versucht, die Gesichtspunkte zu skizzieren, die dabei erwogen werden müssen und für die die Politik die Verantwortung trägt. Zunächst sollte die Frage beantwortet werden, welche Art von Sprachstandserhebung für den angestrebten Zweck angemessen ist und welche Ergebnisse man realistischerweise daraus ableiten kann. Wenn man sich für einen Test entscheidet, dann sollte dieser in jedem Fall den hier beschriebenen Anforderungen genügen. Der Test sollte kontinuierlich überprüft werden, um seine Funktion und Qualität sicherzustellen. Es darf auch nicht vergessen werden, dass das Ergebnis eines Tests entscheidende Konsequenzen für den Teilnehmenden, für größere Gruppen von Teilnehmenden und für die gesamte Gesellschaft haben kann. Bürgerrechte und Menschenrechte der Testteilnehmenden können betroffen sein. Um einen Sprachtest erfolgreich im Kontext von Integration und Einbürgerung einzusetzen, sollten die politischen Entscheidungsträger - nachdem die Entscheidung für einen Test gefallen ist - in einigen wichtigen Fragen mit dem Testanbieter zusammenarbeiten: beispielsweise hinsichtlich der genauen Definition der Kompetenzen, die getestet werden sollen, und bei der Bestimmung der Ressourcen, die für alle Phasen der Testentwicklung und Testdurchführung zur Verfügung stehen. Oberste Priorität sollte immer die Fairness des Verfahrens für die Teilnehmenden haben.