Erstellt von ALTE
im Auftrag des Europarats/Abteilung für Sprachenpolitik

Handbuch zur Entwicklung und Durchführung von Sprachtests

Zur Verwendung mit dem GER

Diese Übersetzung wurde nach Vereinbarung mit dem Europarat erstellt. Die Verantwortung für die Richtigkeit der Übersetzung liegt allein beim Übersetzer.

Herausgegeben von der telc GmbH, Frankfurt am Main, www.telc.net

1. Auflage 2012

Printed in Germany
ISBN: 978-3-86375-093-0
telc Order Number: 5099-B00-010101

Handbuch zur Entwicklung und Durchführung von Sprachtests

Zur Verwendung mit dem GER

Erstellt von ALTE
im Auftrag des Europarats / Abteilung für Sprachenpolitik

Inhalt

Vorwort

Dieses Handbuch ist eine willkommene Ergänzung der Instrumente zur Unterstützung all derjenigen, die den *Gemeinsamen europäischen Referenzrahmen für Sprachen: Lernen, Lehren, Beurteilen* (GER) verwenden. Wir danken der Vereinigung von Sprachprüfungsanbietern in Europa (Association of Language Testers in Europe – ALTE), die vom Europarat mit der Erstellung dieses Dokuments beauftragt wurde und im Geiste ihres dortigen Status als Internationale Nicht-Regierungsorganisation einen wertvollen Beitrag zum erfolgreichen Einsatz des GER leistet.

Der GER soll – zunächst für die Mitgliedsländer des Europarats – allen im Sprachenbereich Tätigen eine gemeinsame Grundlage zur Reflexion und zum Informationsaustausch bieten, seien sie mit der Lehrerausbildung, der Ausarbeitung von Lehrplänen und Vorgaben für den Sprachunterricht oder dem Erstellen von Lehrbüchern und Prüfungen befasst. Der GER stellt für die Nutzer ein beschreibendes Werkzeug dar: Er ermöglicht die Reflexion von Entscheidungen und Verfahrensweisen sowie die angemessene Einordnung und Koordination der Arbeit zum Wohle der Sprachenlernenden im jeweiligen Kontext. Der GER ist also ein flexibles, an einen spezifischen Verwendungskontext anpassbares Werkzeug – ein grundlegender Aspekt, der im System der Kompetenzstufen seinen Ausdruck findet. Dieses kann jeweils angepasst und flexibel ausgelegt werden, um Lern- und Lehrziele sowie Prüfungen zu entwickeln, und findet Anwendung in der Entwicklung der Referenzniveaus für Sprachkompetenz oder *Reference Level Descriptors* (RLDs) für bestimmte Sprachen und Kontexte.

Die beispielhaft formulierten Deskriptoren, die sowohl von muttersprachlichen als auch von nicht-muttersprachlichen Lehrergruppen aus verschiedenen Bildungssektoren mit unterschiedlichen Anforderungen an Sprachausbildung und Lehrerfahrung als transparent, nützlich und relevant angesehen wurden (GER, Kap. 3), erheben nicht den Anspruch, vollständig oder in irgendeiner Hinsicht normativ zu sein. Vielmehr werden die Nutzer aufgefordert, sie an ihren Kontext und ihren Bedarf anzupassen und sie entsprechend zu ergänzen. Dieses Praxis-Handbuch gibt all denjenigen Orientierung, die in diesem Sinne Sprachprüfungen entwickeln, und bezieht sich dabei grundsätzlich auf die GER-Kompetenzstufen, ohne diese jedoch vorschreiben zu wollen.

Die Notwendigkeit, Qualität, Kohärenz und Transparenz beim Lehren von Sprachen sicherzustellen, und das wachsende Interesse an der unbeschränkten Einsetzbarkeit von Qualifikationen hat zur steigenden Bedeutung der GER-Kompetenzstufen und ihrer Nutzung als Referenz- und Messinstrument in Europa und darüber hinaus geführt. Wir freuen uns darüber, ermutigen aber zugleich alle Nutzer zur Erkundung von weiteren Verwendungsmöglichkeiten des GER in seinen zahlreichen Dimensionen und zur Weitergabe ihrer Erfahrungen. Dadurch wird die lebenslange (uneinheitliche und dynamische) Entwicklung eines mehrsprachigen Profils der Sprachlernenden anerkannt und unterstützt, die schließlich die Verantwortung für die Planung und Bewertung ihres Lernfortschritts im Lichte der sich verändernden Umstände übernehmen müssen. Die Initiative des Europarats, mehrsprachige und interkulturelle Bildung zu fördern und hierfür einen globalen Ansatz für alle Sprachen zu entwickeln, führt zu neuen Herausforderungen bei der Entwicklung von Lehrplänen, beim Lehren von Sprachen und nicht zuletzt bei der für die Lernenden genauso wichtigen Bewertung ihrer Sprachkompetenz und der Anwendung ihrer mehrsprachigen und interkulturellen Fähigkeiten. Wir freuen uns auf die unentbehrliche Unterstützung von professionellen Organisationen wie ALTE bei unseren Bemühungen, die Wertevorstellungen des Europarats im Bereich der Sprachbildung zu fördern.

Joseph Sheils
Abteilung für Sprachpolitik, Europarat

Einleitung

Hintergrund

Seit der Veröffentlichung seiner endgültigen Version von 2001 erfährt der *Gemeinsame europäische Referenzrahmen für Sprachen* (GER) ein wachsendes Interesse in Europa und darüber hinaus. Sein Einfluss hat die Erwartungen übertroffen und zweifelsohne dazu beigetragen, das Bewusstsein für fundamentale Fragen beim Lernen, Lehren und Beurteilen von Sprache zu stärken. Darüber hinaus hat der Europarat die Aufstellung eines Instrumentariums (das sog. *Toolkit*) angeregt, das Entscheidern, Lehrern, Prüfungsanbietern und anderen interessierten Gruppen die Verwendung des GER näherbringt und erleichtert.

Wie von Daniel Coste (2007), einem der Verfasser des GER, angemerkt, ist sein Einfluss auf die Bewertung von Sprachkompetenz höchst beachtenswert: Die Prozesse bei der Anbindung von Sprachprüfungen an die GER-Kompetenzstufen haben mehr Aufmerksamkeit auf sich gezogen als andere Aspekte des Referenzrahmens. Heute stehen Prüfungsanbietern und anderen Praktikern mit Interesse an Sprachtests viele nützliche Werkzeuge zur Verfügung, zum Beispiel:

- das *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Europarat, 2009)
- das Beiheft *Reference Supplement to the Manual for Relating Examinations to the CEFR* (Banerjee 2004; Verhelst 2004 a,b,c,d; Kaftandjieva 2004; Eckes 2009)
- Beispielmaterial zur Veranschaulichung der GER-Stufen
- Raster zur Inhaltsanalyse für Material zum Sprechen, Schreiben, Hören und Lesen
- sich stetig weiterentwickelnde Beschreibungen zu den Referenzstufen für Englisch und andere Sprachen

Zusätzlich hat der Europarat Foren ins Leben gerufen, in denen Fachleute Überlegungen zum Handbuch und ihre Erfahrungen beim mit den hier vorgeschlagenen Verlinkungsphasen austauschen konnten (z. B. *Reflections on the use of the Draft Manual for Relating Language Examinations to the CEFR*, Cambridge 2007; ein Seminar im Vorfeld der EALTA-Konferenz, Athen 2008).

Die *Association of Language Testers in Europe (ALTE)* als internationale Nicht-Regierungsorganisation mit Beraterstatus im Europarat hat zu den Werkzeugen des GER-Instrumentariums beigetragen, etwa durch die Bereitstellung des *EAQUALS/ALTE European Language Portfolio (ELP)* und die *ALTE Content Analysis Grids*. Außerdem wurde ALTE durch Dr. Piet van Avermaet im Autorenteam für die Erstellung des *Manual for Relating Language Examinations to the CEFR* repräsentiert. Zusammen mit der Abteilung für Sprachenpolitik des Europarats möchte ALTE die Nutzer des GER-Instrumentariums dazu ermutigen, den GER in ihrem eigenen Kontext zur Umsetzung ihrer eigenen Zielvorgaben anzuwenden.

Ziel dieses Handbuchs

Das oben genannte *Manual for Relating Language Examinations to the CEFR* wurde speziell zum Thema Anbindung von Prüfungen an den Referenzrahmen verfasst und beschreibt zusammen mit seinem Beiheft *Reference Supplement* einen allgemeinen Ansatz sowie eine Reihe von Methoden, u. a. zur Definition von Bestehensgrenzen.

Das vorliegende Handbuch versteht sich als Zusatz zum *Manual for Relating Language Examinations to the CEFR*; es konzentriert sich auf Aspekte der Testentwicklung und Prüfungsdurchführung, die in dem anderen Handbuch nicht behandelt werden. Es ist die überarbeitete Version eines früheren Dokuments des Europarats, das als *Users' Guide for Examiners* (1996) und als deutsche Ausgabe unter *Handreichungen für Testautoren* (2005) bekannt ist und einer von vielen *Users' Guides* war, die vom Europarat als Zusatz zum ersten Entwurf des GER von 1996 / 7 in Auftrag gegeben wurden.

ALTE war für die Erstellung der Originalversion verantwortlich. In den letzten zehn Jahren haben die Entwicklungen in der Validitätstheorie und der vermehrte Gebrauch und Einfluss des GER eine gründliche Aktualisierung des Dokuments notwendig gemacht. ALTE hat die Koordination dieser Überarbeitung 2009 / 2010 gerne übernommen, und viele einzelne ALTE-Mitglieder und -Partner haben dazu beigetragen.

Bei der Überarbeitung hat es sich als sinnvoll erwiesen, sich immer wieder den Ursprung und den Zweck des GER ins Gedächtnis zu rufen, so dass dieses neue Handbuch in Aufbau und Ausrichtung die potentiellen Nutzer anspricht.

Als allgemeiner Referenzrahmen war der GER ursprünglich als „Tool for reflection, communication and empowerment" (Trim 2010) gedacht, also als Werkzeug zur Reflektion, Kommunikation und Stärkung der eigenen Handlungskompetenz. Er wurde erarbeitet, um das allgemeine Verständnis der Bereiche des Sprachenlernens, lehrens und beurteilens zu erleichtern. Zudem bietet er eine umfassende Diskussion zum Fremdsprachenlernen und stellt so eine gemeinsame Sprache zur Verfügung, um sich über alle Aspekte dieses Bereichs zu verständigen. Ebenso bietet er eine Reihe von Referenzniveaus zur Identifizierung des Grads der Sprachbeherrschung von Beinahe-Anfängern (A1) bis zu Fortgeschrittenen (C2) für verschiedene Fertigkeiten und Anwendungsbereiche.

Die genannten Leistungen des GER machen ihn zu einem geeigneten Werkzeug, um verschiedene Herangehensweisen in verschiedensten Kontexten miteinander zu vergleichen – in Europa und darüber hinaus. Allerdings kann der GER als allgemeines Referenzmittel nicht ohne weiteres für jeden Kontext herangezogen werden, ohne dass ihn der Nutzer an die eigenen Gegebenheiten und Zielsetzungen anpasst.

Dies haben die Autoren des GER in der Einleitung klar zum Ausdruck gebracht, indem sie in ihren Hinweisen für Benutzer beispielsweise formulieren: „Wir wollen Praktikern NICHT sagen, was sie tun sollen oder wie sie etwas tun sollen." (Seite 8) – ein Punkt, der mehrmals im Text wiederholt wird. Weitere Informationsmaterialien aus dem GER-Instrumentarium, wie das *Manual for Relating Language Examinations to the CEFR*, haben sich diesem Prinzip ebenfalls unterworfen. Dessen Autoren machen deutlich, dass hier nicht die einzige Methode aufgezeigt wird, nach der Prüfungen an den GER angebunden werden können, und dass keine Institution zu einer solche Verlinkung verpflichtet ist (Seite 1).

Anlässlich eines Forums des Europarats 2007 in Straßburg über den Einsatz des GER bemerkte Coste, wie kontextabhängige Verwendungen – auch als bewusste Eingriffe gesehen – je nach Situation verschiedene Formen annehmen, auf verschiedene Niveaus angewendet werden, unterschiedliche Ziele verfolgen und verschiedene Akteure involvieren können. Coste führte weiter aus, alle diese verschiedenen kontextabhängigen Anwendungen seien legitim und sinnvoll, aber – genauso wie der Referenzrahmen selbst eine Reihe von bereits mitgedachten Optionen anbiete – nutzten einige Anwendungen diesen Rahmen aus,

während andere ihn ausdehnen oder darüber hinaus gingen. Wenn man also Überlegungen zur Verlinkung anstellt, so sollte man sich stets dessen bewusst sein, dass der GER nichts vorschreiben will und dass es keinen Königsweg zur Anbindung einer Prüfung an den GER gibt – immer im Zusammenhang mit dem jeweiligen Kontext und Verwendungszweck gesehen.

Jones und Saville (2009: 54–55) heben hervor:

„… some people speak of applying the CEFR to some context, as a hammer gets applied to a nail. We should speak rather of referring context to the CEFR. The transitivity is the other way round. The argument for an alignment is to be constructed, the basis of comparison to be established. It is the specific context which determines the final meaning of the claim. By engaging with the process in this way, we put the CEFR in its correct place as a point of reference, and also contribute to its future evolution."

(… einige Leute sprechen über die Anwendung des GER auf irgendeinen Kontext wie über die Anwendung von Formeln auf mathematische Probleme. Stattdessen sollten wir aber lieber davon sprechen, einen Kontext auf den GER zu beziehen. Der Bezug ist genau umgekehrt zu sehen. Die Argumentation für eine Anbindung an den GER muss geführt, die Vergleichsgrundlage muss aufgebaut werden. Der spezielle Kontext bestimmt letztlich die Bedeutung der Verlinkungsbehauptung. Wenn wir so verfahren, billigen wir dem GER seine wahre Bedeutung als Referenzrahmen zu und tragen außerdem zu seiner Weiterentwicklung bei.)

Das *Manual for Relating Language Examinations to the CEFR* konzentriert darauf, Vorgehensweisen zum Nachweis des Anspruchs aufzuzeigen, dass sich eine bestimmte Prüfung am GER ausrichtet. Es gibt jedoch keine allgemeinen Leitlinien dazu, wie gute Sprachprüfungen entwickelt werden können, so dass das vorliegende Handbuch einen ergänzenden Ansatz bietet. Es beginnt beim Prozess der Testentwicklung und zeigt, wie die Anbindung an den GER in jedem einzelnen Schritt dieses Prozesses umgesetzt werden kann, um

- den Testinhalt zu bestimmen,
- bestimmte Sprachkompetenzstufen anzuzielen,
- die Leistungen in einem Sprachtest so zu interpretieren, dass sie sich auf die Welt der Sprachverwendung außerhalb des Tests beziehen.

Dieses Handbuch verfolgt also ein weiter gefasstes Ziel als die drei im GER genannten Hauptpunkte, nämlich

- die Spezifikation des Inhalts von Tests und Prüfungen,
- die Kriterien für das Erreichen von Lernzielen, sowohl bei der Beurteilung einer speziellen mündlichen oder schriftlichen Leistung als auch in Bezug auf die kontinuierliche Beurteilung durch Lehrer, andere Lernende oder sich selbst,
- die Beschreibung des Maßes der Sprachbeherrschung in den vorliegenden Tests und Prüfungen, was dann Vergleiche über verschiedene Qualifikationssysteme hinweg ermöglicht.

Es versteht sich als in sich geschlossener Leitfaden für die Testentwicklung im Allgemeinen und somit als nützliches Werkzeug für die Entwicklung von ganz unterschiedlich ausgerichteten Tests. Darüber hinaus stellt es die Testentwicklung in Form eines Zyklus' dar, in dem der erfolgreiche Abschluss einer Phase von der Arbeit in der vorherigen Phase abhängt. Damit jeder einzelne Schritt richtig ausgearbeitet werden kann, muss der gesamte Prozess effizient organisiert werden. Kapitel 1.5 gibt einen Überblick über den Zyklus, der in jedem weiteren Kapitel detailliert ausgearbeitet wird:

- Kapitel 1 führt die grundlegenden Konzepte bei der Prüfung von Sprachbeherrschung ein: Validität, Reliabilität und Fairness.

- Kapitel 2 – „Testentwicklung“ – geht von der Entscheidung, einen Test anzubieten, bis zur Aufstellung der endgültigen Testspezifikationen.
- Kapitel 3 – „Generierung von Testversionen“ – befasst sich mit der Erstellung von Item und Testversionen.
- Kapitel 4 – „Prüfungsdurchführung“ – umfasst die Prüfungsadministration von der Anmeldung der Teilnehmenden bis zur Rücksendung des Materials.
- Kapitel 5 – „Bewertung, Benotung und Übermittlung der Ergebnisse“ – schließt die Durchführungsphase ab.
- Kapitel 6 – „Qualitätssicherung“ – zeigt, wie der Zyklus wiederholt werden kann, um den Test im Laufe der Zeit zu verbessern und immer nützlicher zu machen.

Die Leserinnen und Leser dieses Handbuchs

Dieses Handbuch richtet sich an alle, die an der Entwicklung und Verwendung eines am GER ausgerichteten Sprachtests interessiert sind. Es richtet sich sowohl an neue als auch erfahrene Sprachtestanbieter und zeigt Grundsätze, die sich auf das Prüfen von Sprachbeherrschung im Allgemeinen beziehen – für große Institutionen, die Prüfungen für Tausende von Prüfungsteilnehmerinnen und -teilnehmern an verschiedenen Orten anbieten, genauso wie für die individuelle Lehrkraft, die ihre eigene Lernergruppe prüfen will. Die Grundsätze sind immer die gleichen, für Tests von großer wie von geringerer Bedeutung für die Teilnehmenden, auch wenn die Schritte zur praktischen Umsetzung natürlich variieren.

Wir gehen davon aus, dass die Leserinnen und Leser mit dem GER vertraut sind oder sich bei der Anwendung dieses Handbuchs für die Entwicklung und Durchführung von Tests mit ihm befassen werden.

Der Gebrauch dieses Handbuchs

Auch wenn die hier vorgestellten Prinzipien zum Sprachenprüfen allgemein anwendbar sind, muss der Testanbieter entscheiden, wie er sie in seinem spezifischen Kontext nutzt. Das Handbuch führt Beispiele an und gibt Ratschläge und Anregungen, wie bestimmte Aktivitäten durchgeführt werden können. Dennoch sind diese praktischen Ratschläge in einem Kontext sicherlich passender als in einem anderen, je nach Zweck der Prüfung und den verfügbaren Mitteln für ihre Entwicklung. Das heißt aber nicht, dass dieses Handbuch für bestimmte Leserinnen und Leser weniger sinnvoll ist; wenn die Nutzer die Grundprinzipien verstehen, werden ihnen die Beispiele bei der Überlegung helfen, wie diese Prinzipien an ihren speziellen Kontext angepasst werden können.

Neben dem GER selbst gibt es weitere nützliche Quellen, die zeigen, wie ein Sprachtest an den GER angebunden werden kann. Dieses Handbuch ist nur Teil eines Instrumentariums, bestehend aus Materialien, die vom Europarat entwickelt und verfügbar gemacht wurden. Hier sind deshalb keine Informationen oder theoretischen Ansätze wiedergegeben, die auch an anderer Stelle leicht nachzulesen sind. Vor allem versucht dieses Handbuch, wie oben erwähnt, die Informationen aus dem *Manual for Relating Language Examinations to the CEFR* nicht zu wiederholen, sondern zu ergänzen.

Dieses Handbuch muss nicht im Ganzen gelesen werden. Wenn einzelne Aufgaben bei der Testentwicklung und Prüfungsdurchführung von verschiedenen Personen durchgeführt werden sollen, so kann sich jede Person nur auf den für sie relevanten Teil beschränken. Dennoch kann dieses Handbuch auch für diejenigen, die nur auf einen bestimmten Teil des Testens spezialisiert sind, einen guten Überblick über den gesamten Testentwicklungsprozess geben.

Am Ende jedes Kapitels steht eine Liste mit Schlüsselfragen, die das Verständnis für das Gelesene schärfen sollen. Weiterführende Literaturhinweise geben den Leserinnen und Lesern Hinweise auf ausführlichere Darstellungen oder zu praktischen Arbeitshilfen.

Dieses Handbuch will nichts vorschreiben, sondern die wichtigsten Prinzipien und Ansätze bei der Testentwicklung und -durchführung hervorheben, auf die sich der Nutzer bei der Entwicklung seiner eigenen Tests beziehen kann. Es ist kein Rezeptbuch, um Testaufgaben auf Basis der GER-Beispielskalen zu entwickeln. Denn obwohl die sechs Kompetenzstufen des GER als allgemeines Referenzwerkzeug ausreichend klar und ausführlich sind, wurden sie nicht als Grundlage für eine präzise Gleichsetzung mit Prüfungsleistungen erstellt.

In einem der früheren Entwürfe des Referenzrahmens (Straßburg 1998) wurden die Deskriptorenskalen als Beispiele im Anhang aufgeführt und erschienen nicht im Text selbst. Die einzigen im Haupttext aufgeführten Skalen waren die allgemeinen Referenzstufen. Der ursprüngliche Aufbau des Texts im Entwurf von 1998 betonte die Unterschiede in Rang und Funktion der *allgemeinen Referenzstufen* und der spezifischeren *Deskriptorenskalen*. Dieser Ansatz untermauerte den eher zurückhaltenden, nicht vorschreibenden Ansatz der Skalen, zumal einige nicht kalibriert und diejenigen auf den C-Stufen weniger ausgeführt waren.

In dem Entwurf des GER von 1998 wurde der nicht vorschreibende, vorläufige Status der Deskriptorenskalen explizit im Text erwähnt (Seite 13):

„The establishment of a set of common reference points in no way limits how different sectors in different pedagogic cultures may choose to organise or describe the system of levels and modules. It is also to be expected that the precise formulation of the set of common reference points, the wording of the descriptors, will develop over time as the experience of member states and of institutions with related expertise is incorporated into the description."

(Die Aufstellung eines Sets an gemeinsamen Bezugspunkten begrenzt in keiner Weise, wie verschiedene Bereiche in unterschiedlichen pädagogischen Kulturen ihr System der Niveaus und Module organisieren und beschreiben. Es ist zu erwarten, dass sich der genaue Wortlaut der gemeinsamen Bezugspunkte und der Deskriptoren im Laufe der Zeit weiterentwickelt, da die Erfahrungen der Mitgliedsländer und ähnlich ausgerichteter Fachinstitutionen einfließen werden.)

Würden die Skalen auf allzu präskriptive Weise benutzt, könnten sie als „Patentrezept" für die Beurteilung sprachlicher Leistungen verstanden werden. Die funktionalen und sprachlichen Skalen wollen jedoch zeigen, wie weit gefasst die Sprachniveaus zu verstehen sind, und gerade keine genaue Definition für sie geben. Angesichts der Unterschiede in demografischen Gegebenheiten, Kontexten und Zielen sowie verschiedener Lehr- und Lernmethoden ist es nicht möglich, beispielsweise den „typischen" B1-Lerner zu charakterisieren. Demzufolge ist es auch schwierig, einen Lehrplan oder einen Test für B1 oder jedes andere Sprachniveau zu entwickeln, der für alle Gegebenheiten passt.

Damit der GER eine dauerhafte und positive Wirkung hat, müssen seine Prinzipien und Praktiken in die Routineverfahren der Testanbieter einfließen. Dies wird mit zunehmender Entwicklung der entsprechenden Methoden dazu führen, dass die Anbindung von Tests an den GER tatsächlich argumentativ untermauert werden kann. Dabei kommt der GER als Grundlage der Arbeit zum Einsatz, er muss jedoch je nach spezifischen Kontexten und Anwendungen auch adaptiert werden.

Eine stabile und dauerhafte Ausrichtung auf den GER kann nicht sichergestellt werden, indem einmal bestimmte Standards festgesetzt werden. Vielmehr müssen Testanbieter über einen längeren Zeitraum verschiedene Nachweise hierfür erbringen. Das bedeutet, dass die Empfehlungen aus dem *Manual for Relating Language Examinations to the CEFR* und andere Verfahren aus dem Instrumentarium zur Verlinkung mit dem GER in die Standardverfahren des Testanbieters integriert und nicht als „Einmal-Maßnahme" angesehen werden sollten.

Hierzu möchte dieses Handbuch den Leser ermutigen. Zu betonen ist dabei die Notwendigkeit zur Entwicklung und Aufrechterhaltung von Systemen, die es ermöglichen, Standards zu setzen und diese regelmäßig zu überprüfen.

Im Handbuch verwendete Begriffe

Dieses Handbuch verwendet die nachstehenden Begriffe wie folgt:

Das *Handbuch zur Entwicklung und Durchführung von Sprachtests* wird als „dieses Handbuch" bezeichnet.

Der *Gemeinsame europäische Referenzrahmen für Sprachen: Lernen, Lehren, Beurteilen* wird als „GER" bezeichnet.

Die für die Erstellung eines Tests verantwortliche Einrichtung wird als „Testanbieter" bezeichnet. Ausdrücke wie „Testentwickler" werden gelegentlich benutzt, um diejenigen zu bezeichnen, die innerhalb eines Testentwicklungsprozesses eine bestimmte Funktion innehaben.

Ausdrücke, die im Glossar (Anhang VIII) aufgeführt sind, werden fett gedruckt, wenn sie zum ersten Mal in diesem Handbuch erscheinen oder es hilfreich für den Leser erscheint, diese noch einmal hervorzuheben.

Dr. Michael Milanovic
ALTE Manager

Vorwort zur deutschen Übersetzung

Die deutsche Fassung dieses Handbuchs fußt auf umfangreichen Diskussionen zur Fachterminologie. Viele Begriffe werden in der täglichen Arbeit professioneller Testentwickler auf Englisch verwendet, sollten aber nun möglichst ins Deutsche übertragen werden. Nur so lässt sich sicherstellen, dass eine breitere Fachöffentlichkeit leichten Zugang zur vorliegenden Darstellung der Prozesse zur Testentwicklung und Durchführung von Prüfungen findet.

Wann immer englische Fachbegriffe allgemein geläufig sind – etwa die *Multiple-Choice*-Aufgabe – wurden sie direkt verwendet oder als Alternative neben die deutsche Übersetzung gestellt. Die feinen fachsprachlichen Unterscheidungen beispielsweise zwischen Prüfern, Bewertern und Auswertern werden in den jeweiligen Kapiteln erläutert. Hinweise finden sich zudem im Glossar, wo die wichtigsten Termini erläutert werden.

Im Literaturverzeichnis wurden gegenüber der Originalausgabe einige Titel ergänzt, die für den deutschsprachigen Raum von Bedeutung sind. Beispiele zu Testaufgaben wurden neu gestaltet, so dass auch hier die Rezeption über das Englische vermieden wird.

Dieses Handbuch soll in der deutschsprachigen Fachöffentlichkeit dazu beitragen, die Diskussion zu Grundsätzen des qualitätsgesicherten Testens von fremdsprachlicher Kompetenz voranzubringen. Als Teil des GER-Instrumentariums wird es die Verwendung des GER als Grundlage – nicht aber fertige Anleitung – für Sprachprüfungen verdeutlichen und die deutsche Rezeption des Referenzrahmens verfeinern.

Dr. Sibylle Plassmann
telc – language tests

1 Grundzüge

Ziel dieses Handbuchs ist, praktische Leitlinien zur Entwicklung von Sprachtests vorzustellen. Dazu ist eine theoretische Verankerung notwendig, die im ersten Kapitel anhand der folgenden Punkte dargelegt wird:

- Definition von Sprachbeherrschung
- Validität als grundlegendes Charakteristikum eines zweckmäßigen Tests
- Reliabilität
- Fairness

Außerdem werden die Grundzüge des Testentwicklungsprozesses aufgezeigt und in den weiteren Kapiteln näher beschrieben.

1.1 Zur Definition von Sprachbeherrschung

1.1.1 Modelle der Sprachverwendung und der Sprachkompetenz

Sprachverwendung ist ein hoch komplexes Phänomen, das zahlreiche verschiedene Fertigkeiten und Kompetenzen abruft. Bei der Entwicklung eines Sprachtests ist es wichtig, von einem klar formulierten Modell dieser verschiedenen Fertigkeiten auszugehen und aufzuzeigen, wie diese miteinander in Verbindung stehen. Ein solches Modell kann, muss aber nicht zwingendermaßen den Anspruch erheben, genau aufzuzeigen, wie Sprachkompetenz in unserem Gehirn tatsächlich angelegt ist. Es soll in erster Linie die entscheidenden Aspekte der Sprachbeherrschung aufzeigen, die wir zu beachten haben. Es dient als Ausgangspunkt für die Entscheidung, welche Aspekte von Sprachverwendung oder – kompetenz getestet werden können und sollten, und es hilft sicherzustellen, dass die Testergebnisse interpretierbar und verwendbar sind. Die in dem Modell aufgezeigte mentale Eigenschaft wird auch als **Eigenschaft** oder **Konstrukt** bezeichnet.

1.1.2 Das GER-Modell zur Sprachverwendung

Maßgebliche Modelle der Sprachkompetenz wurden von zahlreichen Autoren vorgeschlagen (z. B. Bachman 1990, Canale und Swain 1981, Weir 2005).

Für dieses Handbuch ist es sinnvoll, mit dem *Gemeinsamen europäischen Referenzrahmen für Sprachen: Lernen, lehren, beurteilen* (GER) zu beginnen, der ein allgemeines Modell zur Sprachverwendung und zum Erlernen einer Sprache vorschlägt. Dieser **handlungsorientierte Ansatz** wird in einem Abschnitt wie folgt erklärt:

„… umfasst die Handlungen von Menschen, die als Individuen und als gesellschaftlich Handelnde eine Vielzahl von *Kompetenzen* entwickeln, und zwar *allgemeine*, besonders aber *kommunikative Sprachkompetenzen*. Sie greifen in verschiedenen *Kontexten* und unter verschiedenen *Bedingungen und Beschränkungen* auf diese Kompetenzen zurück, wenn sie sprachliche *Aktivitäten* ausführen, an denen (wiederum) *Sprachprozesse* beteiligt sind, um *Texte* über bestimmte *Themen* aus verschiedenen *Lebensbereichen* (Domänen) zu produzieren und/oder zu rezipieren. Dabei setzen sie *Strategien* ein, die für die Ausführung dieser *Aufgaben* am geeignetsten erscheinen. Die Erfahrungen, die Teilnehmer in solchen kommunikativen Aktivitäten machen, können zur Verstärkung oder zur Veränderung der Kompetenzen führen." (GER o. S., Hervorhebungen im Original).

Dieser Abschnitt identifiziert die Hauptelemente des Modells, welche später im Text des GER näher vor-

gestellt werden. Tatsächlich können die Überschriften und Zwischenüberschriften der Kapitel 4 und 5 des GER als ein hierarchisches Modell von Elementen gelesen werden, die in umfassendere Elemente eingebettet sind.

Abbildung 1 illustriert dies anhand einiger Überschriften und Zwischenüberschriften aus Kapitel 5, „Die Kompetenzen des Sprachverwendenden/Lernenden". Sie zeigt zwei Seiten von Kompetenz: Allgemeine Kompetenzen (wie z.B. Deklaratives Wissen und Weltwissen, hier nicht gezeigt) und Kommunikative Sprachkompetenzen, die in drei weitere Kompetenzen unterteilt sind: Linguistische, Soziolinguistische und Pragmatische Kompetenzen. Jede ist wiederum weiter unterteilt.

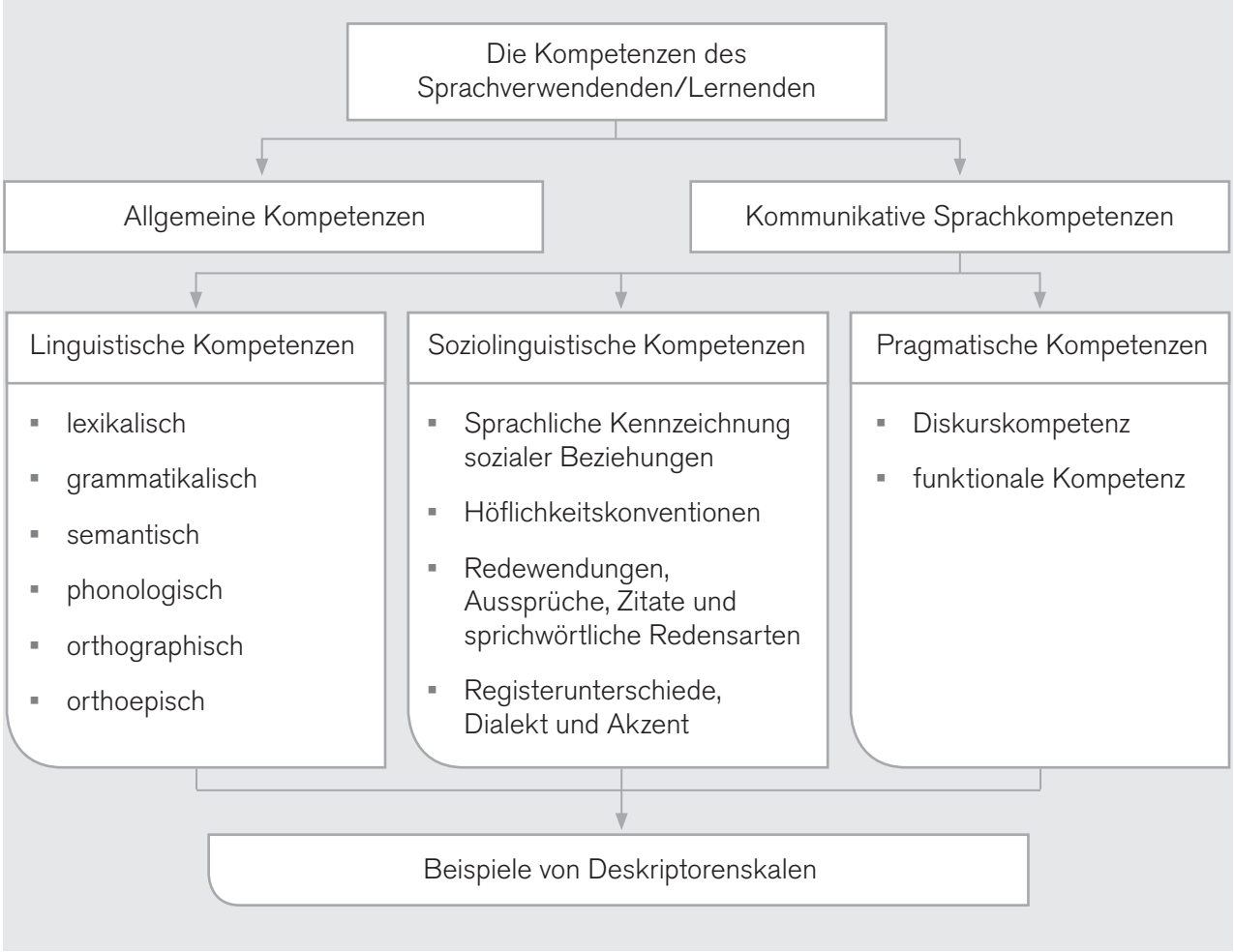


Abb. 1: Eine Teildarstellung des GER, Kapitel 5: Die Kompetenzen des Sprachverwendenden/Lernenden

In ähnlicher Weise beschreibt Kapitel 4 die kommunikativen Absichten und die Art, in der Sprache verwendet wird. Wie in Abbildung 2 dargestellt, enthält es Überlegungen bezüglich der Frage, *was* Kommunikation ist (Themen, Aufgaben und Absichten), aber auch bezüglich der Aktivitäten und Strategien und demzufolge der funktionalen Sprachfähigkeiten, die die Lernerinnen und Lerner beim Kommunizieren zeigen. Der Übersichtlichkeit wegen zeigt Abbildung 2 nur einen Teil der komplexen Hierarchie.

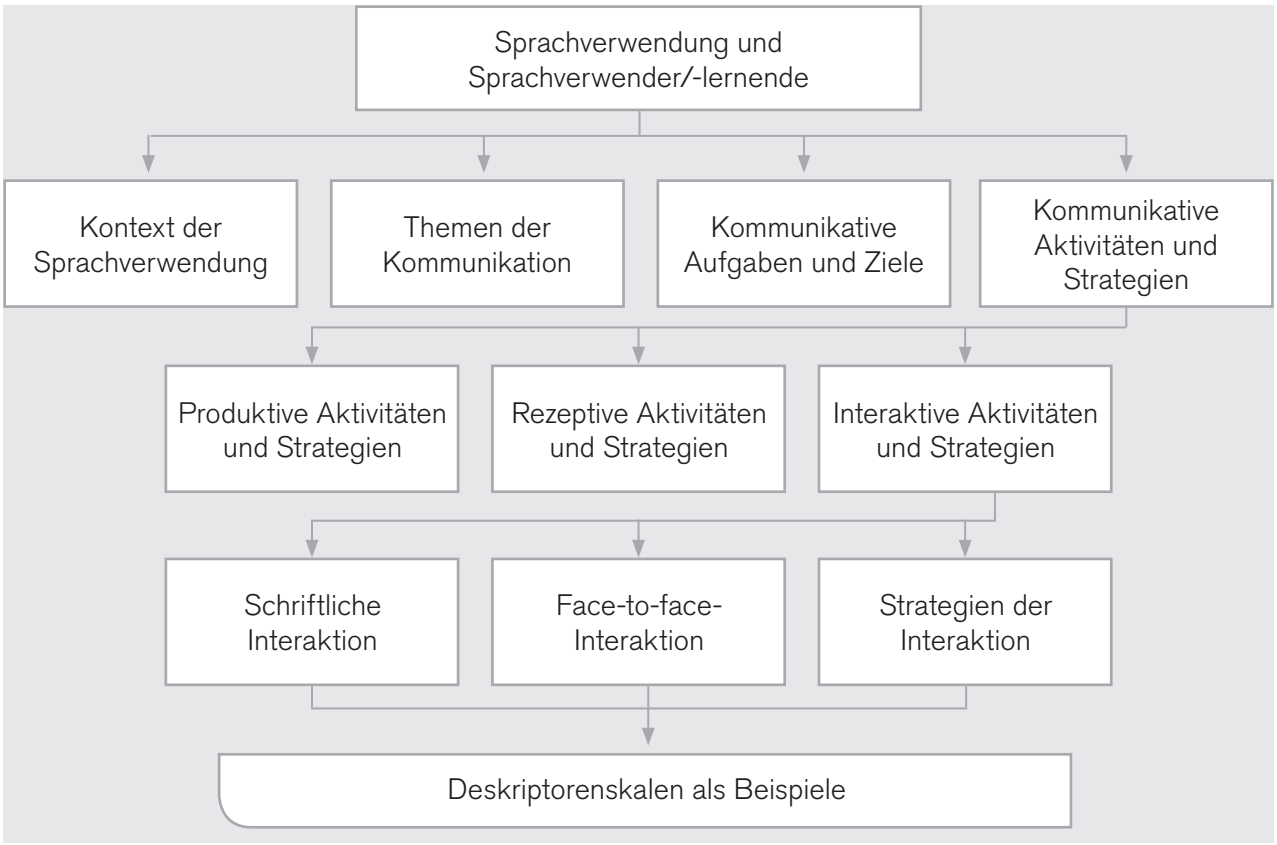


Abb. 2: Eine Teildarstellung des GER, Kapitel 4: Sprachverwendung, Sprachverwender und Sprachenlernende

1.1.3 Umsetzung des Modells

Bei der Frage, wie das **Modell zur Sprachverwendung** praktisch umgesetzt werden kann, müssen zwei wichtige Aspekte berücksichtigt werden, die beträchtlichen Einfluss auf das Testformat haben: Die **Authentizität** der **Items** und **Aufgaben** sowie die Frage, inwieweit die Kompetenzen unabhängig voneinander getestet werden.

Authentizität

Die zwei wichtigen Aspekte bei der Authentizität von Sprachtests sind die *situative* und die *interaktionale* Authentizität. *Situative* Authentizität beschreibt die Genauigkeit, mit der Aufgaben und Items Sprachaktivitäten aus dem tatsächlichen Leben wiedergeben. *Interaktionale* Authentizität dagegen beschreibt die Natürlichkeit der Interaktion zwischen den Teilnehmenden und der Aufgabe und den zugrunde liegenden mentalen Prozess. Eine Testaufgabe zum Hörverstehen einer spezifischen Situation zum Beispiel kann situativ authentischer gestaltet werden, wenn eine Sprachverwendungssituation aus dem täglichen Leben nachgestellt wird, z. B. der Wetterbericht. Wird der Prüfungsteilnehmerin oder dem -teilnehmer eine Motivation zum Zuhören gegeben, z. B. die Planung eines Picknicks in dieser Woche und die Auswahl eines bestimmten Tages, so ist die Aufgabe in der Interaktion authentischer.

In Sprachtests müssen oft verschiedene Aspekte der Authentizität abgewogen werden, um eine angemessene Aufgabe zu erstellen. So ist es notwendig, Materialien und Aktivitäten an das jeweilige Sprachniveau des Lerners in der Zielsprache anzupassen. Diese Anpassung bedeutet, dass die Situation, auf die sich

Lernende einlassen, und die Interaktion mit den Texten und anderen Lernenden aber durchaus authentisch sein *kann* – auch wenn das Material selbst sprachlich nicht authentisch ist.

Um ein Item oder eine Aufgabe situativ authentischer zu gestalten, müssen die wesentlichen Charakteristika der jeweiligen Sprachhandlung im täglichen Leben identifiziert und so weit wie möglich nachgebildet werden. Mehr interaktionale Authentizität kann wie folgt erreicht werden:

- Situationen und Aufgaben verwenden, die den Teilnehmenden wahrscheinlich vertraut sind und mit denen sie auf dem gegebenen Sprachniveau zu tun haben
- den Zweck und die intendierten Adressaten der jeweiligen Aufgabe verdeutlichen, indem man eine angemessene Kontextualisierung erzeugt
- Kriterien für das erfolgreiche Lösen der Aufgabe deutlich machen

Kompetenzen miteinander verbinden

Bei der Definition eines Modells zur Sprachverwendung mag es scheinen, als ob die einzelnen Kompetenzen für sich alleine stünden. Es ist jedoch sehr schwierig, in einer authentischen Aufgabe die einzelnen Kompetenzen klar voneinander zu trennen, da ein kommunikativer Akt immer viele Kompetenzen gleichzeitig erfordert. Wenn zum Beispiel ein Sprachlernender versucht, jemanden zu verstehen, der ihn auf der Straße angehalten hat, um nach dem Weg zu fragen, kommen mehrere Kompetenzen zum Tragen: Kompetenz in Grammatik und Textverständnis, um die Nachricht zu dekodieren; soziolinguistische Kompetenz, um den gesellschaftlichen Kontext zu verstehen, in dem die Kommunikation stattfindet; und illokutive Kompetenz, um zu verstehen, welchen Zweck der Sprecher verfolgt.

Beim Entwurf einer Testaufgabe ist es wichtig, die Kompetenzen klar abzuwägen, die für die erfolgreiche **Lösung** der **Aufgabe** benötigt werden. Einige Kompetenzen werden wichtiger sein als andere – diese stehen dann im Mittelpunkt der Aufgabe. Die Aufgabe sollte sprachlich so gestaltet sein, dass die Fertigkeit der Teilnehmenden in der ausgewählten Kompetenz oder den ausgewählten Kompetenzen beurteilt werden kann. Die Art, wie die Antwort **ausgewertet** oder **bewertet** wird, muss ebenfalls berücksichtigt werden (siehe Kapitel 2.5 und 5.1.3): Die Fertigkeit in der/den ausgewählten Kompetenz(en) sollte die Grundlage der Bewertung sein.

1.1.4 Die Kompetenzstufen des GER

Zusätzlich zu dem oben aufgeführten Modell stellt der GER einen Referenzrahmen von sechs kommunikativen Sprachniveaus als Hilfe bereit, um Lernziele zu setzen und den Lernfortschritt oder die Sprachbeherrschung messbar zu machen. Dieser konzeptionelle Rahmen wird durch Skalen von **Deskriptoren** veranschaulicht, die in Form von Kann-Beschreibungen dargestellt werden.

Ein Beispiel für eine Kann-Beschreibung für die niedrigste Stufe (A1), Leseverstehen, lautet wie folgt:
Kann vertraute Namen, Wörter und ganz einfache Sätze verstehen, z. B. auf Schildern, Plakaten oder in Katalogen.

Im Vergleich dazu der Deskriptor der höchsten Stufe (C2):
Kann praktisch jede Art von geschriebenen Texten mühelos lesen, auch wenn sie abstrakt oder inhaltlich und sprachlich komplex sind, z. B. Handbücher, Fachartikel und literarische Werke.

Die sechs Kompetenzstufen lauten wie folgt:

C2 Mastery C1 Effective Operational Proficiency	} kompetente Sprachverwendung
B2 Vantage B1 Threshold	} selbstständige Sprachverwendung
A2 Waystage A1 Breakthrough	} elementare Sprachverwendung

Als Sprachtester sollten wir die Kann-Beschreibungen genau verstehen. Diese sind

- veranschaulichend.

Daher sind sie:

- nicht erschöpfend,
- nicht präskriptiv

und

- keine Definition,
- kein Lehrplan,
- keine Checkliste.

Die Kann-Beschreibungen dienen als Leitfaden für den Bildungsbereich, so dass man Kompetenzstufen erkennen und über sie sprechen kann. Wir können die Kann-Beschreibungen als Leitfaden für die Testentwicklung verwenden, aber ihre bloße Übernahme entbindet nicht von der Aufgabe, Kompetenzstufen für den Test genauer festzulegen.

Testentwickler müssen entscheiden, welche der Kann-Beschreibungen in ihrem Kontext von größter Relevanz sind. Zum Beispiel die Sprachverwendungsbereiche (**Domänen**) des Tests: Wird Hotelpersonal unterrichtet und getestet, sind die Deskriptoren für „praktische zielorientierte Zusammenarbeit" (GER: 4.4.3.1) sicherlich nützlich, diejenigen zum Thema „Fernsehsendungen und Filme verstehen" (GER: 4.4.2.3) dagegen wahrscheinlich nicht. Wenn die im GER oder in Zusatzmaterialien zur Verfügung stehenden illustrativen **Skalen** den Prüfungskontext nicht ausreichend abdecken, so können sie durch Kann-Beschreibungen aus anderen Quellen oder für diesen Kontext neu geschriebene Deskriptoren ergänzt werden.

Die Anbindung von Tests an den GER

Wir sehen also, dass die **Verlinkung** von Tests mit dem GER mit der Anpassung desselben an den entsprechenden Testkontext beginnt. Denn der GER ist darauf angelegt, „kontextfrei (zu) sein, um Raum zu lassen für generalisierbare Ergebnisse aus verschiedenen spezifischen Kontexten". Gleichzeitig jedoch soll der GER „kontextrelevant sein, also auf alle nur denkbaren relevanten Kontexte bezogen und in sie übersetzt werden können." (GER: 24).

Die Verlinkung der Tests sollte nicht der Versuch sein, den GER in jedem möglichen Kontext starr und mechanisch anzuwenden. Testentwickler müssen die Art und Weise rechtfertigen können, wie sie den GER auf ihren Kontext beziehen oder übertragen, unter anderem durch Erläuterung der Besonderheiten ihres speziellen Kontextes.

Auch die Merkmale der Prüfungsteilnehmerinnen und teilnehmer sind kontextuelle Faktoren. So gibt es große Unterschiede zwischen Lernenden hinsichtlich ihres Alters und ihrer kognitiven Entwicklung, des Zwecks des Spracherwerbs usw. Einige dieser Unterschiede definieren geradezu bestimmte Lernergrup-

pen. Sprachtests werden oft eigens für eine spezielle Lernergruppe entwickelt, z. B. für Jugendliche oder Erwachsene. Beide Gruppen können in Bezug zum GER gesetzt werden, aber junge Lernende auf der GER-Stufe B1 werden eine andere Ausprägung von B1-Merkmalen zeigen als Erwachsene auf der Stufe B1, weil andere Deskriptoren zum Tragen kommen.

Lernende unterscheiden sich häufig auch in ihrem Kompetenzprofil (einige können besser hörend verstehen als lesend, bei anderen ist es umgekehrt). Das macht es schwer, sie mithilfe einer einzigen Skala zu vergleichen. So ist es also möglich, dass zwei Teilnehmende aufgrund unterschiedlicher Stärken und Schwächen beide auf der Stufe B1 eingeordnet werden. Wenn eine differenzierte Bewertung unterschiedlicher Fertigkeitsbereiche wichtig ist, sollten diese Fertigkeiten getrennt geprüft werden und spezielle Deskriptoren als Grundlage für die Definition von Kompetenzstufen in den einzelnen Fertigkeiten dienen.

Eine entscheidende Einschränkung muss allerdings bei der Anpassung des GER an einen bestimmten Kontext beachtet werden. Der GER dient nur zur Beschreibung von Sprachkompetenz gemäß dem in Kapitel 1.1.2 dieses Handbuchs dargestellten Modells zur Sprachverwendung. Man sollte nicht versuchen, außerhalb dieses Modells liegende Kenntnisse oder Kompetenzen hiermit zu verbinden, wie z. B. das Verstehen von fremdsprachlicher Literatur.

1.2 Validität

1.2.1 Was ist Validität?

Validität kann ganz einfach definiert werden als das Ausmaß, in dem ein Test das misst, was er messen soll. Wenn unser Test beispielsweise die kommunikativen Fähigkeiten in Italienisch testen soll und die Teilnehmerinnen und Teilnehmer systematisch gute oder schlechte Ergebnisse je nach ihren Fertigkeiten im Italienischen erzielen, dann ist unser Test valide. Diese ziemlich enge Definition wurde in den letzten Jahren erweitert, um auch die Art und Weise mit einzubeziehen, wie Tests *verwendet* werden. Validität bezieht sich also auf das Ausmaß, in dem die Interpretation von Prüfungsergebnissen im vorgesehenen Verwendungskontext durch praktische und theoretische Belege unterstützt wird.

Diese weiter gefasste Definition betont die gesellschaftliche **Wirkung** von Prüfungen und die Notwendigkeit, angemessene Informationen im Zusammenhang mit möglicherweise sehr wichtigen Entscheidungen über die Geprüften bereitzustellen. So können wir also nicht von einer absoluten Validität sprechen, sondern müssen festhalten, dass Tests nur so weit valide sind, wie die Testergebnisse zu einem bestimmten Zweck verwendet werden. Es ist die Interpretation des Testergebnisses für den einzelnen Teilnehmer, die valide oder nicht valide ist.

Bachman (1990) bezieht dies speziell auf den Bereich „Sprache“, indem er Rückschlüsse auf einen Bereich des Zielsprachengebrauchs aufgrund des Testergebnisses einfordert. Das heißt, um die Validität der Testergebnisse beurteilen zu können, müssen wir zunächst festlegen, was Prüfungsteilnehmerinnen oder -teilnehmer bei der Sprachverwendung in einer tatsächlichen Situation können müssen, und dann entscheiden, ob der Test einen guten Nachweis dieser Kompetenzen erlaubt. Der GER liefert einen wertvollen Ansatz, um den Erfolg in bestimmten Verwendungsbereichen darzustellen. Seine veranschaulichenden Deskriptoren dienen dabei als Ausgangspunkt.

1.2.2 Validität und der GER

Wenn wir Testergebnisse mit Hilfe des GER übermitteln, so behaupten wir, Prüfungsleistungen in Bezug auf unsere Definition von Teilnehmerinnen und Teilnehmern auf einer bestimmten GER-Stufe interpretieren zu können. Validität bedeutet in dem Zusammenhang, dass unsere Behauptung wahr ist: dass wir für unsere Behauptung, ein Lerner befinde sich auf der GER-Stufe B1, tatsächlich entsprechende Nachweise erbringen können.

Die Art der benötigten Nachweise variiert je nach Kontext der Prüfung. Das GER-Modell für den Sprachverwendenden/Lernenden wie oben erklärt kann als *sozio-kognitiv* bezeichnet werden: Sprache ist sowohl ein verinnerlichtes Repertoire von Kompetenzen als auch ein nach außen gewandtes Repertoire sozialer Verhaltensweisen. Je nach Kontext bezieht sich ein Sprachtest mehr auf das eine oder mehr auf das andere, was wiederum Einfluss auf den Nachweis der Testvalidität hat.

Liegt der Fokus auf Sprachverwendung, so bezieht sich die Darlegung der Validität auf die für verschiedene kommunikative Zwecke tatsächlich verwendete Sprache.

Liegt der Fokus eher auf Kompetenz, lässt sich die Validität in Bezug auf kognitive Fertigkeiten, Strategien und Sprachwissen belegen, was einen Rückschluss auf die potentiellen Fähigkeiten zur Sprachverwendung zulässt.

Im zweiten Fall ist es wichtig zu zeigen, dass die Testaufgaben kognitive Fertigkeiten, Strategien und Bereiche des Sprachwissens erfordern, die in der Verwendungssituation der Zielsprachen notwendig sind – es muss also *interaktionale Authentizität* herrschen (siehe 1.1.3).

Beide Arten von Belegen können die im GER beschriebene Validität von Sprachtests unterstützen. Die Abwägung zwischen ihnen hängt von den Anforderungen des spezifischen Kontextes ab. In einem Sprachtest für Verkaufspersonal wäre vermutlich die Fähigkeit zur Sprachverwendung sehr wichtig, wohingegen ein Sprachtest für Schulkinder mehr Gewicht auf Kompetenz legen könnte.

1.2.3 Validität im Prozess der Testentwicklung

Validität verbindet also die Leistung bei einer Testaufgabe mit dem Rückschluss auf die Sprachfertigkeit der Prüfungsteilnehmerin oder des -teilnehmers in einer Situation außerhalb der Prüfung. Es leuchtet ein, dass die Erstellung einer Testaufgabe ein entscheidender Schritt in diesem Zusammenhang ist; andere Schritte sind jedoch ebenso wichtig.

In diesem Abschnitt wird Validität auf den Prozess einer Testerstellung im engeren Sinn bezogen (siehe 1.5), so dass der Einfluss auf andere Phasen in der Testentwicklung deutlich wird. Die Phasen der Testerstellung meinen eine Reihe von aufeinander folgenden Schritten, wobei jeder einzelne Schritt erfolgreich abgeschlossen sein muss, wenn der abschließende Rückschluss über die Teilnehmenden valide sein soll.

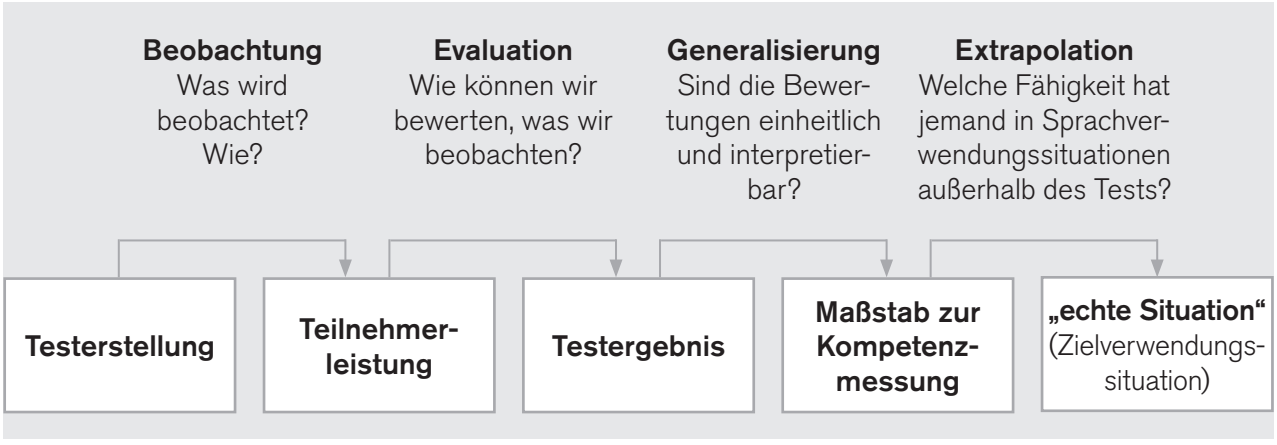


Abb. 3: Argumentationskette in einer Validitätsdiskussion (nach Kane, Crooks und Cohen 1999, Bachman 2005)

Abbildung 3 stellt diese Schritte dar:

1. Der Test wird so entwickelt, dass eine Teilnehmerleistung herauskommt, die auf einem Modell der Lernerkompetenzen basiert und interpretierbar ist. Beispielsweise kann ein Teilnehmer gebeten werden, einen Brief an einen Freund zu einem bestimmten Thema zu schreiben.

2. Die Teilnehmerleistung wird bewertet. Welche Aspekte der Teilnehmerleistung werden belohnt, welche führen zu Abwertung? In unserem Beispiel beziehen sich diese Aspekte auf die kommunikativen Fähigkeiten wie im Modell zur Sprachverwendung beschrieben, inklusive Register (soziolinguistische Kompetenz), lexikalische, grammatische und orthographische Kompetenz (linguistische Kompetenz) etc.

3. Bis hierher besteht die Testbewertung aus Zahlen, die sich ausschließlich auf eine einzige Leistung bei einer bestimmten Aufgabe beziehen. Wie können diese generalisiert werden – würden die Teilnehmenden das gleiche Resultat bei einer anderen Gelegenheit oder bei einer anderen Testversion erzielen? Diese Frage führt zur Reliabilität (siehe Kapitel 1.3). Ein weiterer Aspekt der Generalisierung betrifft die Abbildung auf eine weiter gefasste Kompetenzskala, da z. B. eine Testversion einfacher sein kann als eine andere und wir dies feststellen und ggf. ausgleichen wollen (siehe Anhang VII).

4. Bis jetzt haben wir die Leistung innerhalb der „Welt“ des Tests beschrieben, aber wir möchten auf die Welt außerhalb des Tests schließen. Hier stellen wir die Verbindung zu den GER-Stufen her, indem wir – geleitet durch Kann-Beschreibungen – die Sprachkompetenz der Teilnehmenden in tatsächlichen Verwendungssituationen benennen.

5. Auf dieser Grundlage können wir Entscheidungen über die geprüfte Person treffen.

Diese kurze Auflistung verdeutlicht, dass Validität, inklusive einer angestrebten Verbindung zum GER, von jedem einzelnen Schritt in der Testerstellung und -durchführung abhängt. Validität ist Bestandteil des gesamten Prozesses.

Anhang I gibt Hinweise zum Aufbau einer Validitätsargumentation.

1.3 Reliabilität

1.3.1 Was ist Reliabilität?

Reliabilität beim Testen bedeutet Konsistenz: Ein Test mit reliabler Bewertung führt bei jedem Einsatz zu gleichen oder ähnlichen Resultaten. Dies bedeutet, dass ein Test eine Gruppe von Teilnehmenden immer in nahezu die gleiche Rangfolge bringen würde. Es bedeutet *nicht* zwingend, dass jeweils dieselben Teilnehmerinnen und Teilnehmer den Test bestehen oder nicht bestehen würden, denn die Bestehensgrenze kann anders angesetzt sein. Daher wird von Abhängigkeit (Dependenz) gesprochen, wenn wir sowohl an der Konsistenz als auch an der Genauigkeit des Prüfungsergebnisses interessiert sind.

Es gilt zu beachten, dass hohe Reliabilität nicht unbedingt heißen muss, dass der Test gut oder die Interpretation der Ergebnisse valide ist. Ein schlechter Test kann sehr reliable Ergebnisse erzielen. Aber auch das Gegenteil trifft nicht unbedingt zu: Für eine valide Interpretation der Testergebnisse muss die Auswertung eine akzeptable Reliabilität aufweisen, sonst sind die Ergebnisse weder zuverlässig noch aussagekräftig.

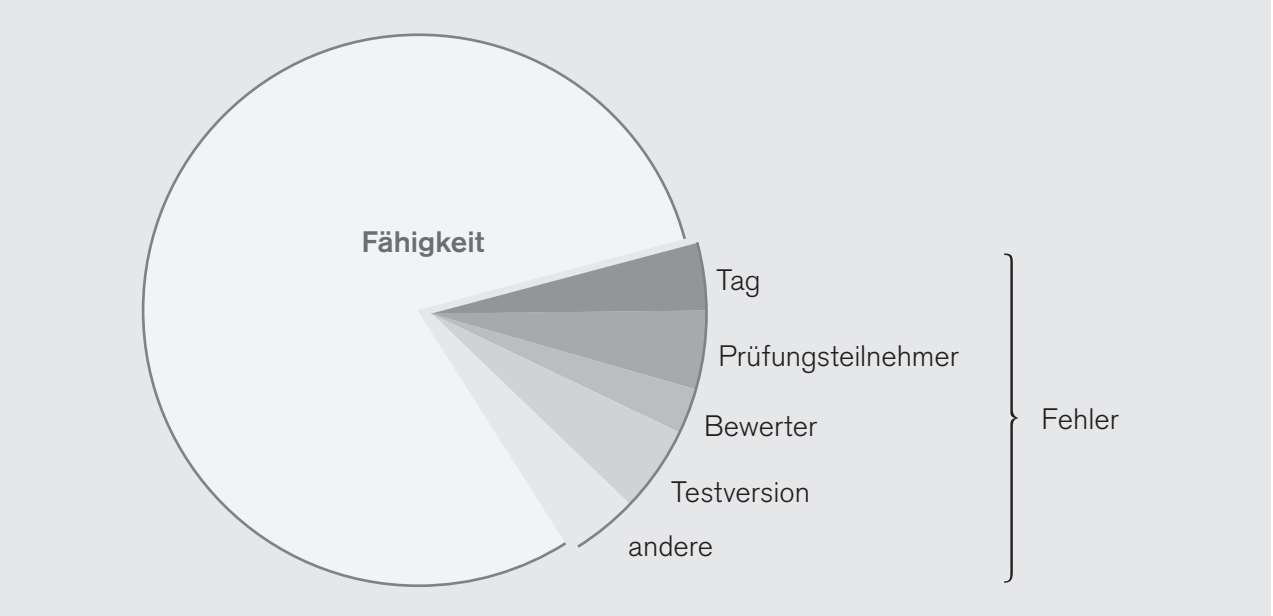


Abb. 4: Einige Ursachen für Fehler bei der Testauswertung

Prüfungsergebnisse variieren von Teilnehmer zu Teilnehmer. Reliabilität wird definiert als jener Anteil an der Varianz der Testergebnisse, der durch die gemessene Fähigkeit begründet wird und nicht durch andere Faktoren. Die durch andere Faktoren begründete Variabilität wird als **Messfehler** bezeichnet. Es gilt zu beachten, dass das Wort *Fehler* nicht im Sinne von Nachlässigkeit benutzt wird. Jeder Test enthält ein bestimmtes Maß an Fehlern.

Abbildung 4 zeigt einige verbreitete Ursachen für Messfehler:

- der Tag der Prüfung (unterschiedliche Wetterbedingungen, andere Prüfungsdurchführung usw.)
- bessere oder schlechtere Leistungsfähigkeit des jeweiligen Prüfungsteilnehmers am jeweiligen Tag
- unterschiedliche Leistungen der Bewerterinnen und Bewerter oder Varianz aufgrund der jeweiligen Testversion
- sonstige nicht kontrollierbare Faktoren

Unser Ziel ist es, Tests zu entwickeln, bei denen der Anteil der Ergebnisvariabilität aufgrund von Teilnehmerfähigkeit den Anteil der Variabilität durch Fehler deutlich übertrifft.

1.3.2 Reliabilität in der Praxis

Der Testentwickler sollte sich der möglichen Ursachen von Messfehlern bewusst sein und diese möglichst minimieren. Das Befolgen der in diesem Handbuch beschriebenen Verfahrensweisen und Grundsätze kann dazu beitragen. Die statistische Analyse zur Einschätzung der Reliabilität von Testergebnissen ist zudem ein wichtiger *post hoc*-Schritt nach dem Einsatz des Tests. Anhang VII enthält weitere Informationen über Verfahrensweisen der Reliabilitätseinschätzung.

Reliabilitätsziele für die Ergebnisse jeglicher Tests kann es nicht geben, da Reliabilitätswerte davon abhängen, wie sehr die Ergebnisse der Teilnehmenden variieren. Ein Test für eine Gruppe von Lernern, die bereits ein Auswahlverfahren durchlaufen haben, wird typischerweise geringere Reliabilitätswerte aufweisen als ein Test für eine sehr uneinheitliche Lernerpopulation. Die Beurteilung der Reliabilität hängt auch von den Items oder den Aufgabentypen und der Art der Auswertung ab. Nach Kriterien bewertete Aufgaben (siehe

Kapitel 5) sind typischerweise weniger reliabel als **dichotome Items**, da bei einer kriterienorientierten Bewertung größere Abweichungen (Messfehler) auftreten können als bei einer Auswertung nach festem Lösungsschlüssel.

In jedem Fall ist es empfehlenswert, die Überprüfung der Reliabilität routinemäßig durchzuführen. So lässt sich herausfinden, welche Tests besser oder schlechter funktionieren, und die verbesserte Qualität von Tests über die Zeit kann kontrolliert werden. Die meisten Reliabilitäts-Schätzwerte wie etwa Cronbachs Alpha oder KR-20 liegen in der **Spannweite** von 0 bis 1. Um eine Faustformel zu geben: Werte im oberen Drittel der Spanne (0.6 bis 1) werden in den meisten Fällen als akzeptabel angesehen.

Eine statistische Reliabilitätseinschätzung ist jedoch gewöhnlich nicht möglich, wenn die Anzahl der Teilnehmenden und/oder der Items gering ist. In diesen Fällen ist nicht festzustellen, ob die Reliabilität für den Zweck des Tests angemessen ist oder nicht. Eine gute Strategie zur Kompetenzfeststellung ist in solchen Fälle die Verwendung des betreffenden Tests als nur eine der Entscheidungsgrundlagen. Zusätzliche Informationen können aus einem Portfolio mit Arbeitsproben, aus einer Reihe von Tests über einen längeren Zeitraum hinweg oder aus anderen Quellen bezogen werden.

1.4 Ethische Standards und Fairness

1.4.1 Gesellschaftliche Auswirkungen des Prüfens

Messick (1989) hat sich für die entscheidende Rolle von Werten und **Wirkung** einer Prüfung als Teil der Validität ausgesprochen. Sein Einfluss hat das Augenmerk mehr auf den gesellschaftlichen Wert von Prüfungen und ihre Auswirkungen auf alle **Beteiligten** gelenkt. Zu den Folgen und zur Wirkung von Prüfungen gehören beabsichtigte (und hoffentlich positive) Ergebnisse des Prüfens, aber auch die ungeahnten und manchmal negativen Nebeneffekte, die Prüfungen haben können. Die Einführung einer neuen Prüfung kann beispielsweise die Art, wie Lehrerinnen und Lehrer unterrichten (positiv oder negativ) beeinflussen („Washback-Effekt").

Testanbieter sollten Untersuchungen zu diesem Rückwirkungseffekt und anderen Auswirkungen anstellen, um mehr über die gesellschaftlichen Folgen ihrer Prüfungen zu erfahren. Solche Untersuchungen können auch schon in einem sehr kleinen Rahmen stattfinden. In der Unterrichtspraxis lässt sich zum Beispiel feststellen, ob Lernende wegen einer bestimmten Ausrichtung des Tests vielleicht einigen Aspekten des Lernstoffes mehr Aufmerksamkeit schenken und dafür andere vernachlässigen. Ein Methodenwechsel zur verstärkten Arbeit an den vernachlässigten Aspekten kann notwendig sein, ggf. auch ein veränderter Testfokus.

1.4.2 Fairness

Ein Ziel für alle Test- und Prüfungsanbieter ist es, ihre Tests so fair wie möglich zu machen. Als Referenz hierfür gelten der *Code of Fair Testing Practices in Education* (JCTP 1988) und die *Standards for Educational and Psychological Testing* (AERA et al 1999).

Die 1999 erstellten *Standards* identifizieren drei Aspekte der Fairness: Fairness durch Vermeidung von Verzerrungsfaktoren (Bias), Fairness als Gleichbehandlung im Prüfungsprozess und Fairness als Gleichheit bei den Testergebnissen.

Kunnans „Test Fairness Framework" (Kunnan 2000a, 2000b, 2004, 2008) konzentriert sich auf fünf Aspekte, die bei Sprachprüfungen berücksichtigt werden müssen, um Fairness zu erreichen: Validität (siehe Kapitel 1.2), die Vermeidung von verzerrenden Faktoren bei der Bewertung (siehe Anhang VII), Zugänglichkeit, Durchführung (siehe Kapitel 4) und gesellschaftliche Auswirkungen.

Verschiedene Institutionen haben einen Verhaltens – oder Fairness-Kodex aufgestellt, um Testanbieter in den praktischen Aspekten zur Sicherstellung von Fairness zu unterstützen.

Testanbieter versuchen in der Regel, Voreingenommenheit schon bei der Testentwicklung zu minimieren. Beispielsweise können bestimmte Themen (z. B. lokale Bräuche) bestimmten Gruppen von Teilnehmenden Vor- oder Nachteile bringen (z. B. denjenigen aus Ländern mit ganz anderen Bräuchen). Eine Liste solcher in Tests zu vermeidenden Themen wird den Testautoren zur Verfügung gestellt. Zu beachtende Gruppen von Prüfungsteilnehmerinnen und -teilnehmern definiert man unter anderem durch Alter, Geschlecht oder Nationalität, wobei dies vom jeweiligen Prüfungskontext abhängt (siehe 3.4).

1.4.3 Ethische Bedenken

Seit den frühen 80er Jahren werden auch ethische Bedenken in Bezug auf Sprachprüfungen diskutiert. Vor allem Spolsky (1981) warnte vor den negativen Folgen, die besonders wichtige (*High Stakes*-) Prüfungen für die Teilnehmenden haben können, und sprach sich dafür aus, Prüfungen wie Medikamente zu etikettieren: „Mit Vorsicht einzusetzen". Er konzentrierte sich dabei besonders auf Sprachtests für bestimmte Zwecke, z. B. im Zusammenhang mit Migration, bei denen die Entscheidungen über eine Person aufgrund des Prüfungsergebnisses ernste und weitreichende Folgen haben können.

Die *International Language Testing Association (ILTA)* veröffentlichte im Jahr 2000 ihren *Code of Ethics*; dieser legt umfassende Richtlinien zum Verhalten von Testanbietern dar.

Testanbieter müssen sicherstellen, dass sich alle Beteiligten innerhalb ihrer Einrichtung der maßgeblichen Prinzipien bewusst sind und dass sie diese genau verstehen. Nur so kann man sicherstellen, dass diese Leitlinien tatsächlich eingehalten werden. Außerdem sind weitere Maßnahmen zur Sicherstellung von Fairness angebracht (siehe Kapitel 4 und Anhang VII).

1.5 Arbeitsschritte

Die einzelnen Phasen der Testentwicklung und -durchführung bilden einen Zyklus, bei dem der Erfolg jeder Phase von den Ergebnissen der vorangegangenen Phase abhängt. Daher ist es wichtig, den gesamten Prozess gut zu steuern. Dabei ist das Sammeln von Belegen von entscheidender Bedeutung, um wichtige Entscheidungen fundiert treffen zu können.

Abbildung 5 zeigt die Phasen bei der Entwicklung eines neuen Tests. Der Zyklus beginnt mit der Entscheidung, eine Prüfung zur Verfügung zu stellen. Diese kommt entweder vom Testanbieter selbst oder von anderer Stelle, etwa von einem Schulleiter, einer Verwaltungsstelle oder einem Ministerium. Darauf folgt die Phase der Testentwicklung im engeren Sinne, gefolgt von der Prüfungsdurchführung. Jede dieser Phasen enthält zahlreiche kleinere Arbeitsschritte, die jeweils abgeschlossen werden müsen, um die Phase als Ganzes abschließen zu können. Welches Ziel mit jedem Schritt des Prozesses erreicht werden soll, ist auf der rechten Seite des Diagramms dargestellt. Eine Zeitachse verdeutlicht wie die einzelnen Phasen und Arbeitsschritte aufeinander aufbauen, denn der Beginn eines neuen Arbeitsschritts erfordert immer die erfolgreiche Beendigung des vorhergehenden. Wurde das Testformat einmal entwickelt, können die verschiedenen Schritte der Prüfungsdurchführung regelmäßig wiederholt werden. Diesem Durchführungszyklus liegt das einmal erarbeitete Ergebnis der Testentwicklungsphase, nämlich die **Testspezifikationen,** zugrunde. So generiert man immer wieder **Testversionen** der gleichen Prüfung.

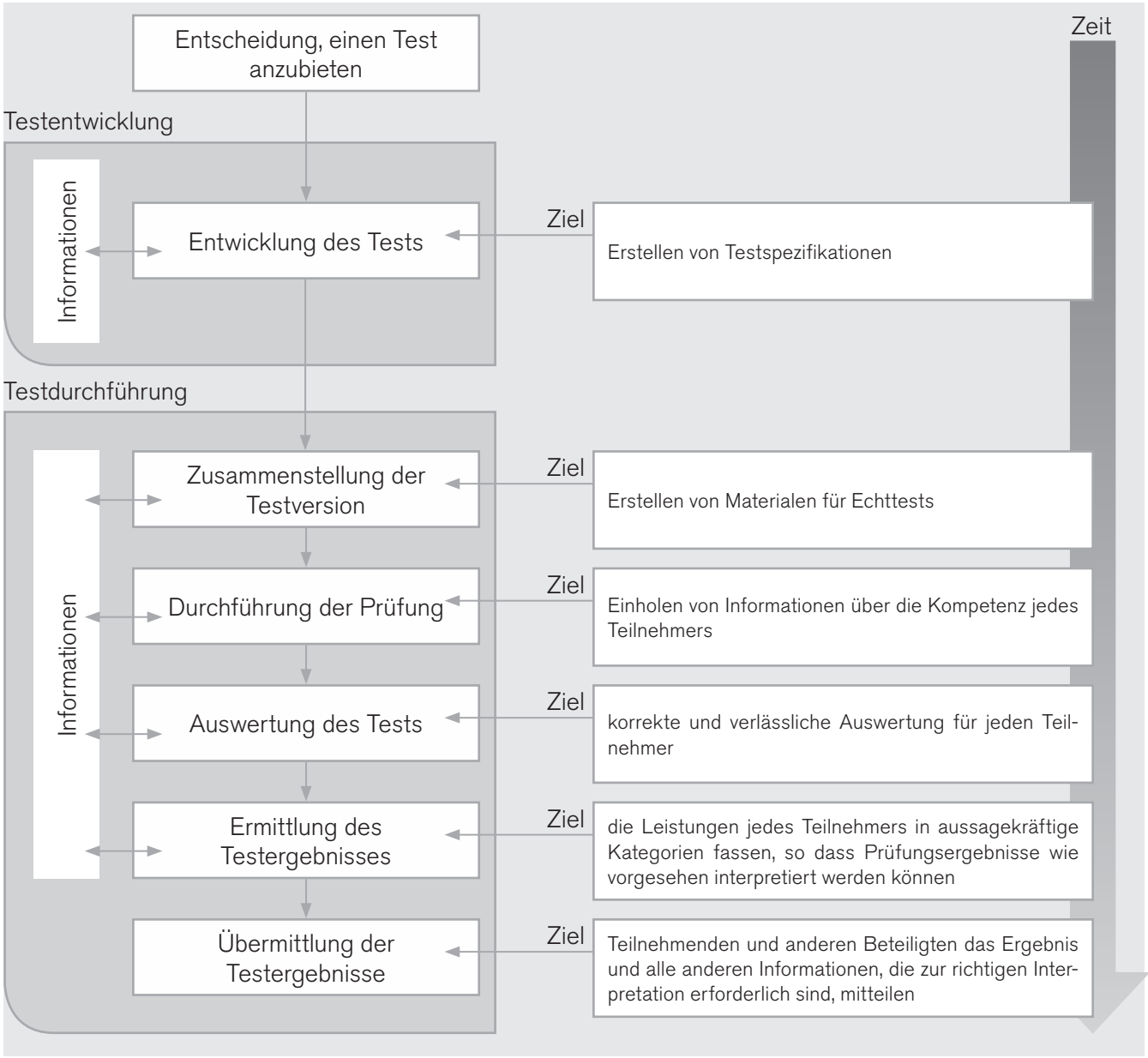


Abb. 5: Der allgemeine Prozess der Testentwicklung und -durchführung

Die in Abbildung 5 aufgeführten Phasen finden bei jeder Testentwicklung Anwendung, unabhängig davon, wie groß oder klein die Institution des Testanbieters ist.

Jede dieser Phasen in Abbildung 5 setzt sich aus kleineren, so genannten Mikro-Aufgaben und Aktivitäten zusammen. Diese werden in späteren Kapiteln dieses Handbuchs genauer beschrieben. Jeder Arbeitsschritt sollte standardisiert sein, um sicherzustellen, dass jede neu generierte Testversion in hohem Maße vergleichbar mit vorangegangenen Versionen ist.

Die Sammlung und Verwendung von Belegen wird in den Kästen auf der linken Seite des Diagramms dargestellt. Nachweise, wie z. B. Hintergrundwissen über die Prüfungsteilnehmerinnen und -teilnehmer, Rückmeldungen von anderen Beteiligten, die Teilnehmerleistungen in Bezug auf Aufgaben und Items sowie die Zeit, in der die Teilnehmenden bestimmte Aufgaben bewältigen, sind ein wichtiger Kontrollfaktor für die Testentwicklung und später für den Nachweis, dass die empfohlene Verwendung der Testergebnisse valide ist.

Solche Belege sollten routinemäßig gesammelt und verwendet werden, da dieser wichtige Aspekt ansonsten im Prozess der Testentwicklung leicht vergessen wird.

1.6 Schlüsselfragen

- Welche Aspekte des GER-Modells zur Sprachverwendung sind für Ihren Kontext geeignet?
- Welche GER-Kompetenzstufen passen am besten?
- Wie sollen Ihre Testergebnisse verstanden und interpretiert werden?
- Was könnte die Reliabilität in Ihrem Kontext am meisten gefährden?
- Wie können Sie sicherstellen, dass Ihre Arbeit sowohl ethisch als auch fair gegenüber den Prüfungsteilnehmerinnen und -teilnehmern ist?
- Welche Herausforderungen können in der Planung Ihrer Testentwicklung auftreten?

1.7 Weiterführende Literatur

Modelle der Sprachverwendung

Fulcher und Davidson (2007: 36–51) diskutieren Konstrukte und Modelle näher.

Validität

ALTE (2005: 19) gibt eine nützliche Zusammenfassung der Validitätstypen und des Hintergrunds der modernen Konzeption von Validität.

Kane (2004, 2006), Mislevy, Steinberg und Almond (2003) erörtern verschiedene Aspekte des Validitätsnachweises (die auch in Anhang I dieses Handbuchs aufgeführt werden) und machen genauere Angaben zu deren Ausführung.

Reliabilität

Traub und Rowley (1991) sowie Frisbie (1988) beschreiben die Reliabilität von Testergebnissen auf leicht verständliche Weise. Parkes (2007) zeigt, wann und wie die Informationen aus einem einzelnen Test durch andere Informationen ergänzt werden können, um Entscheidungen über die geprüfte Person zu treffen.

Ethische Standards und Fairness

Seit den frühen 1990er Jahren wurden spezielle *Codes of Practice* für Prüfungsteilnehmende von professionellen Sprachtestverbänden entwickelt, z. B.

- Der *ALTE Code of Practice* (1994)
- Die *EALTA Guidelines for Good Practice in Language Testing and Assessment* (2006)
- Die *ILTA Guidelines for Practice* (2007)

In den 1990er Jahren erschien eine Sonderausgabe von *Language Testing* mit Alan Davies als Gastherausgeber (1997), die sich auf ethische Standards in Sprachprüfungen konzentriert. 2002 wurde in Passadena eine Konferenz zu „Language assessment ethics" abgehalten. Die Berichte zu dieser Konferenz sind in einer Sonderausgabe von *Language Assessment Quarterly* zusammengefasst (ebenfalls mit Gastautor Davies, 2004). McNamara und Roever (2006) geben einen Überblick über Fairness-Kontrollen und die Ethikrichtlinien für Prüfungen.

Mehrere Artikel in *Language Testing*, April 2010 (Davies 2010, Kane 2010, Xi 2010) stellen dar, wie man Belege für Testfairness sammelt und in Form eines umfassenden Nachweises zusammenstellt.

2 Testentwicklung

2.1 Der Prozess der Testentwicklung

Ziel bei der Entwicklung eines neuen Tests ist es, **Testspezifikationen** zu produzieren, die für die Erstellung von **Echttests** genutzt werden können. Testentwicklung beginnt mit der Entscheidung einer Person oder einer Institution (dem Testauftraggeber), dass ein neuer Test gebraucht wird. Abbildung 6 zeigt den Testentwicklungsprozess und seine drei wichtigsten Phasen (Planung, Formatentwicklung und Erprobung) sowie eine weitere Phase (Information der Beteiligten), die in einigen Kontexten zum Tragen kommt, auch wenn sie nicht wie die anderen Arbeitsschritte zur Produktion von Testspezifikationen beiträgt, sondern lediglich zur Information über den neuen Test dient.

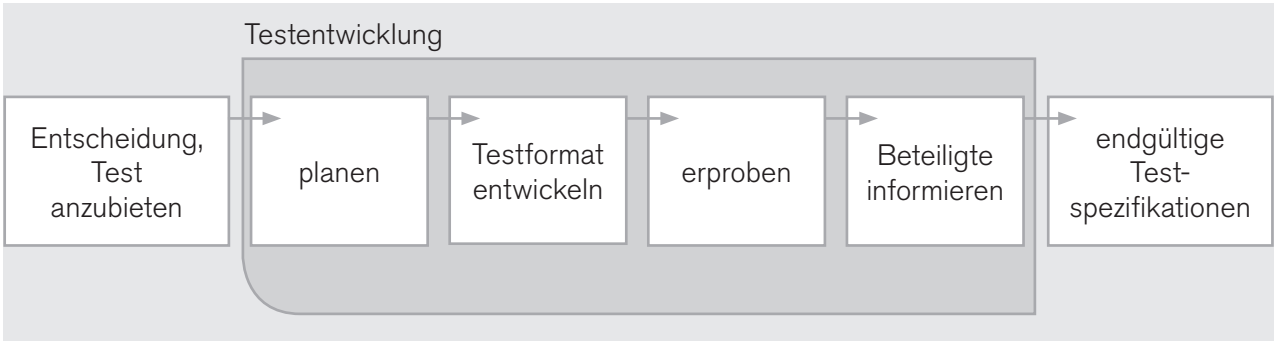


Abb. 6: Der Prozess der Testentwicklung

2.2 Die Entscheidung, einen Test anzubieten

Diese Entscheidung wird hier nicht als Teil des Testentwicklungsprozesses angesehen. Sie ist jedoch von ausschlaggebender Bedeutung für die Planungsphase, da die Anforderungen des Auftraggebers einen entscheidenden Einfluss auf das Testformat und dessen Verwendung haben.

Wer entscheidet, dass eine neue Sprachprüfung gebraucht wird? In einigen Fällen wird diese Entscheidung auf eigene Verantwortung vom Testanbieter getroffen. In anderen Fällen hat ein Auftraggeber bereits entschieden, dass eine neue Prüfung gebraucht wird, und tritt von außen an den Testanbieter heran.

In beiden Fällen müssen die Anforderungen klar identifizierbar sein, was zusätzliche Arbeit für die Testentwickler bedeuten kann. Es ist ggf. schwierig, die Ziele eines Testauftraggebers zu verstehen, da dieser nicht zu derselben Einrichtung gehört oder er aufgrund mangelnder Expertise im Testen oder Unterrichten von Sprache nicht weiß, welche Informationen der Testentwickler benötigt.

2.3 Planung

Diese Phase ist dem Sammeln von Informationen gewidmet, die man für die späteren Stufen der Testentwicklung benötigt. Viele dieser Informationen sollten vom Auftraggeber kommen. Jedoch kann auch eine größere Gruppe von Beteiligten konsultiert werden, wie z. B. Ministerien und andere staatliche Stellen, Verlage, Schulen, Eltern, Fachleute, Arbeitgeber, Bildungsinstitute und Prüfungszentren. Wenn viele solche Ansprechpartner involviert sind, können die Informationen mit Hilfe von Fragebögen oder in Fachseminaren erhoben werden. Im Unterricht dagegen ist das persönliche Wissen über den Kontext und die Lernenden häufig ausreichend. Die wichtigsten Fragen, die sich ein Testentwickler stellen muss, lauten wie folgt:

- Welche Merkmale haben die Prüfungsteilnehmerinnen und -teilnehmer?
  (Alter, Geschlecht, sozialer Hintergrund, Bildungshintergrund, Erstsprache etc.)
- Welchen Zweck soll die Prüfung erfüllen?
  (Schulabschlusszeugnis, Zulassung zu einem Bildungsangebot, Mindestanforderungen für einen Beruf, Unterstützung des Lernfortschritts, Lernstandsdiagnose etc.)
- In welchem Bezug steht der Test zu einem Bildungskontext?
  (Lehrplan, methodischer Ansatz, Lernziele etc.)
- Welcher Standard wird für den vorgeschlagenen Zweck benötigt?
  (eine GER – Kompetenzstufe in einer bestimmten Fertigkeit, ein Standard für einen speziellen Verwendungsbereich etc.)
- Wie werden die Prüfungsergebnisse verwendet?

Die Antworten auf diese Fragen ermöglichen es dem Entwickler, mit der Definition der zu testenden Sprachkompetenz zu beginnen und festzulegen, wie die Bestehensgrenze gesetzt werden muss, um Entscheidungen über die Teilnehmenden zu treffen (siehe Kapitel 5), und wie das Testergebnis den Beteiligten dargestellt und erläutert werden soll (siehe Kapitel 5).

Fragen zu den Auswirkungen des Tests in einem weiteren Kontext sind ebenfalls hilfreich:

- Wer sind die Beteiligten?
- Welche Wirkung ist erwünscht?
- Welche Wirkung wird erwartet?

Und schließlich sollten auch eher praktische Fragen nicht vergessen werden:

- Wie viele Prüfungsteilnehmende werden erwartet?
- Wann sollte der Test bereitstehen?
- Wie wird die Prüfung finanziert und wie groß ist das Budget?
- Wie oft wird die Prüfung durchgeführt?
- Wo wird die Prüfung durchgeführt?
- Wie soll der Test angeboten werden (auf Papier oder computergestützt)?
- Wer ist verantwortlich für die einzelnen Phasen der Prüfungsverwaltung (d. h. Produktion des Materials und Generierung von Testversionen, Durchführung, Auswertung, Übermittlung der Ergebnisse)?
- Welche Auswirkungen haben die Durchführungsmodalitäten auf die Prüfungssicherheit? Sollen z. B. eine oder mehrere Testversionen zum Einsatz kommen?
- Welche Auswirkungen wird dies auf die Logistik haben, d. h. wird die Arbeit des Testanbieters von anderen Stellen, wie z. B. Testzentren, abhängen?
- Wie wird die Testqualität langfristig sichergestellt?
- Wird eine Erprobung möglich und durchführbar sein (siehe Kapitel 3.4)?

2.4 Formatentwicklung

Die Informationen aus der Planungsphase dienen als Ausgangspunkt für die Entwicklungsphase. Die wichtigen Entscheidungen über die Art des Tests werden getroffen und die ersten Testspezifikationen werden erarbeitet. Diese beschreiben die allgemeine Struktur des Tests und alle Aspekte seines Inhalts. Detailliertere Testspezifikationen, z. B. für Testautoren und für das mit der Organisation und der Durchführung der Prüfung betraute Personal, können entwickelt werden, wenn die ersten Testspezifikationen verabschiedet wurden.

2.4.1 Ausgangsüberlegungen

Die erste Herausforderung in der Entwicklungsphase besteht darin, eine genauere Vorstellung von Testinhalt und -format zu entwickeln. Den Ausgangspunkt hierfür bilden die Informationen, die zu den Testanforderungen und zum Testhintergrund gesammelt wurden, wie z. B. die Eigenschaften der Teilnehmenden, der Zweck des Tests und die erforderliche Kompetenzstufe.

Der GER ist eine nützliche Quelle, um die Eigenschaften des Tests zu definieren, da viele seiner Kapitel sich direkt auf das Prüfen beziehen, vor allem:

- Kapitel 6 – „Sprachen lernen und Fremdsprachenlehren“ – regt Überlegungen hinsichtlich Lernzielen und Lehrmethodologie an, was wiederum Auswirkungen auf den Stil, den Inhalt und die Funktion des Tests hat.
- Kapitel 7 – „Die Rolle kommunikativer Aufgaben beim Fremdsprachenlernen und -lehren“ – gibt auch Hinweise zur Verwendung von Aufgaben im Test.
- Kapitel 9 – „Beurteilen und Bewerten“ – beschreibt, wie der GER für verschiedene Testzwecke genutzt werden kann.

Sicherlich von größter Relevanz sind jedoch die Kapitel 4 und 5, da diese vom Testinhalt und den zu prüfenden Fertigkeiten handeln. Sie bieten dem Testentwickler viele Möglichkeiten, aus dem insgesamt handlungsorientierten Ansatz des GER und seinem Modell zur Sprachverwendung (Kapitel 1.1) auszuwählen, z. B. mit Bezug auf folgende Aspekte:

- den Fokus der Aufgaben, z. B. das detaillierte Verstehen eines Texts zu zeigen (GER Kapitel 4.4 und 4.5)
- was geprüft werden soll, also Fertigkeiten, Kompetenzen und Strategien (GER Kapitel 5)
- verwendete Textsorten als Input (GER Kapitel 4.6)
- Textquellen (GER Kapitel 4.1 und 4.6)
- einige Hinweise zu Themenbereichen, die für Tests geeignet sind (GER Kapitel 4.1 und 4.2)
- Arten der Vorgaben (*Prompts*) für die mündlichen Prüfungen (GER Kapitel 4.3 und 4.4)
- Beispiele für alltägliche Situationen, die für die Prüfungsteilnehmenden relevant sind (GER Kapitel 4.1 und 4.3)
- die für diese Situationen notwendige Stufe der Sprachbeherrschung (zahlreiche Skalen mit Kann-Beschreibungen im GER)
- Beurteilungskriterien für freie Schreibaufgaben und für mündliche Prüfungen (die jeweiligen Kann-Beschreibungen im GER, z. B. Seite 61)

Zusätzlich muss der Testanbieter einige praktische Fragen des Testens klären, beispielsweise die folgenden:

- Der Umfang der Prüfung:
  In der Regel sollten die Teilnehmenden ausreichend Zeit haben, um alle Testfragen vollständig zu beantworten, ohne sich beeilen zu müssen. Es ist besonders wichtig, dass sie die Möglichkeit haben, ihre tatsächlichen Fähigkeiten unter Beweis zu stellen. Die Bearbeitungszeit muss zunächst von erfahrenen Sprachtestern geschätzt werden, wobei auch andere Testformate zu Rate gezogen werden können (siehe weiterführende Literatur, Kapitel 2.8). Nach der Erprobung des Tests oder auch nach dem ersten Einsatz als Echttest wird die Zeitvorgabe ggf. korrigiert. In einigen Fällen werden auch so genannte *Speed tests* eingesetzt, also Tests, bei denen die Teilnehmenden dazu angehalten werden, zügig zu arbeiten und die Aufgaben in sehr kurzer Zeit zu lösen. Auch in diesem Fall muss die Zeitvorgabe erprobt werden.

- Die Anzahl der Items pro Test:
  Man benötigt eine ausreichende Anzahl von Items, um den gewünschten Inhalt des Tests abzudecken und um genügend reliable Informationen über die Kompetenz der Teilnehmenden zu erhalten. Andererseits gibt es für den Umfang des Tests praktische Beschränkungen.

- Die Anzahl der Items pro Subtest:
  Wenn der Test darauf abzielt, mehrere Aspekte reliabel zu prüfen, so erfordert dies eine ausreichende Anzahl von Items pro Subtest. Man kann dazu andere Testformate zu Rate ziehen und die Reliabilität ermitteln (siehe Anhang VII).

- Die Aufgabentypen:
  Offene und geschlossene Aufgabenformate erfordern unterschiedliche Arbeitsweisen. Zu den geschlossenen Aufgaben gehören *Multiple-Choice*-Aufgaben (Mehrfachwahlaufgaben), *Matching*-Aufgaben (**Zuordnungsaufgaben**) oder *Ordering*-Aufgaben (Erstellung einer Reihenfolge). Offene Aufgabentypen erfordern kurze Antworten (Lückenaufgaben) oder ausführlicheres Schreiben. Die verschiedenen Typen haben Vor- und Nachteile. Siehe ALTE (2005: 111–34) für weitere Informationen über Aufgabentypen.

- Die Gesamt- und Einzellänge der Texte, gemessen z. B. an der Wortzahl:
  Bestehende Testformate (siehe weiterführende Literatur, Kapitel 2.8) können Anhaltspunkte zu praktikablen Textlängen geben.

- Das Testformat:
  Ein Test mit diskreten, also voneinander **unabhängigen Items** besteht aus kurzen Items, die nicht miteinander in Verbindung stehen. In einem aufgabenbasierten Test dagegen sind die Items in kleineren Aufgabengruppen zusammengefasst, die sich z. B. auf einen Lese- oder Hörtext beziehen. Da aufgabenbasierte Testformate längere und authentischere Impulse geben können, sind sie im Allgemeinen angemessener für kommunikative Sprachprüfungen. Siehe ALTE (2005: 134–47) für mehr Informationen über Aufgabentypen.

- Die Anzahl der Punkte pro Item, pro Aufgabe oder **Testteil**:
  Je mehr Punkte pro Item oder Abschnitt möglich sind, desto größer ist dessen jeweilige Gewichtung. Generell ist die beste Lösung, einen Punkt pro Item zu vergeben. Es kann aber auch Gründe geben, die Items verschieden zu gewichten, indem man ihnen mehr oder weniger als einen Punkt zuschreibt (siehe Anhang VII).

- Die Eigenschaften der Bewertungskriterien:
  Werden aufgabenspezifische Kriterien verwendet? Wie umfangreich soll jede Bewertungsskala sein? Ist die Bewertung analytisch oder holistisch? (Kapitel 2.5 und 5.1.3 gehen näher auf Bewertungsskalen ein.)

Die Phase der Formatentwicklung mündet also in grundsätzliche Überlegungen zum Zweck des Tests, den Fertigkeiten sowie den inhaltlichen Bereichen, die dieser abdecken soll, und gibt Angaben zur praktischen Umsetzung. Man sollte ebenfalls Überlegungen anstellen zur Punktevergabe, zu Bewertungskriterien für die produktiven Fertigkeiten Schreiben und Sprechen (siehe Kapitel 2.5), zur Durchführung der Prüfung (Kapitel 4) und dazu, wie **Auswerter** und **Bewerter** ausgebildet und geführt werden können (Kapitel 5.1.3). Alle Beteiligten sollten die diesbezüglichen Vorschläge genau prüfen und auswerten.

Weiterhin muss die Kommunikation mit den Prüfungsteilnehmerinnen und -teilnehmern sowie anderen Beteiligten zu folgenden Punkten bedacht werden:

- wie viele Stunden Vorbereitungszeit (bei zwingender Prüfungsvorbereitung) notwendig sind,
- auf welche Weise Übungstests zur Verfügung gestellt werden,
- welche Informationen den Testnutzern (allen relevanten Beteiligten) vor und nach dem Test gegeben werden.

Schließlich sind auch die Erwartungen der Testnutzer zu bedenken:

- Wie passt der Test in den aktuellen Lehrplan und zur Unterrichtspraxis?
- Was für eine Art Test erwarten die Beteiligten?

Kapitel 4 des GER stellt ein besonders nützliches Referenzsystem zur Verfügung, mit dem alle besonderen Merkmale eines Tests in der Entwicklungsphase abgeglichen werden können. Hierfür wird eine zusammenfassende Darstellung des Testformats in Diagrammform erarbeitet. Dieser Ansatz wird in Anhang III dieses Handbuchs beispielhaft dargestellt. Der Modelltest ist für Teilnehmende auf der GER-Stufe C1 gedacht, die die Sprache im Hochschulkontext lernen, und besteht aus vier Prüfungsteilen. Der Anhang gibt einen Überblick über den gesamten Inhalt der Prüfung sowie eine allgemeine Beschreibung für jeden einzelnen Prüfungsteil.

2.4.2 Berücksichtigung der Durchführungspraxis

An dieser Stelle der Testentwicklung müssen das Testkonzept und praktische Gegebenheiten der Durchführung in Einklang gebracht werden. Informationen zur Durchführung werden zusammen mit den Testanforderungen in der Planungsphase gesammelt (Kapitel 2.3). Der **Testentwickler** muss Anforderungen und Sachzwänge gegeneinander abwägen und seine Entscheidungen vom Auftraggeber genehmigen lassen. Bachmann und Palmer (1996: Kap. 2) geben hierfür mit ihrem Konzept zur **Zweckmäßigkeit** eines Tests einen Rahmen vor. Nach ihrem Ansatz entsteht Zweckmäßigkeit als Zusammenspiel von sechs Merkmalen:

- **Validität**: Die Interpretation von Testergebnissen ist aussagekräftig und angemessen.
- **Reliabilität**: Die erzielten Testergebnisse sind gleich bleibend und beständig.
- **Authentizität**: Die Aufgaben entsprechen echten Sprachverwendungssituationen in den jeweiligen Handlungsfeldern.
- **Interaktivität**: Die Aufgaben erfordern mentale Prozesse und Strategien, die auch in Situationen des täglichen Lebens notwendig sind.
- **Wirkung**: Hier geht es um die – hoffentlich positiven – Auswirkungen, die der Test auf den Einzelnen, auf die Unterrichtspraxis und in der Gesellschaft hat.
- **Praktikabilität**: Die Entwicklung, Erstellung und Durchführung des Tests muss wie geplant und mit den verfügbaren Ressourcen möglich sein.

Diese Merkmale stehen oft im Wettbewerb zueinander: So kann etwa ein höheres Maß an Aufgabenauthentizität zu geringerer Reliabilität führen. Bemühungen zur Steigerung der Zweckmäßigkeit des Tests müssen daher auf die bestmögliche Abwägung aller genannten Aspekte abzielen.

2.4.3 Testspezifikationen

Das Ergebnis der Formatentwicklung ist ein Satz fertiger Testspezifikationen. Der erste Entwurf dieser Spezifikationen enthält Entscheidungen über die meisten der bisher angesprochenen Aspekte. Nach der Erprobung (siehe Kapitel 2.5) werden die Spezifikationen endgültig festgelegt. Je bedeutsamer der Test für die Beteiligten ist, desto bedeutsamer sind auch die Testspezifikationen. Sie stellen die Qualität des Tests sicher und zeigen die Validität der empfohlenen Interpretation der Testergebnisse.

Testspezifikationen sind aber auch wichtig, wenn der Test keine entscheidenden Auswirkungen für seine Teilnehmerinnen und Teilnehmer hat. Sie garantieren, dass die einzelnen Testversionen immer auf derselben Grundlage basieren und dass der Test zum Lehrplan oder zu anderen Aspekten des Prüfungskontexts passt.

Testspezifikationen können auf verschiedene Weise geschrieben werden, je nach Bedarf des Testanbieters und der Zielgruppe. Es wurde eine Reihe von Modellen für Testspezifikationen entwickelt (siehe weiterführende Literatur, Kapitel 2.8); diese können als hilfreicher Ausgangspunkt dienen.

2.5 Pilotierung

Ziel dieser Phase ist es, die erste Fassung der Testspezifikationen in der Praxis zu prüfen und Verbesserungen auf der Grundlage von Erfahrungswerten und Vorschlägen der Beteiligten anzubringen.

Sobald ein Entwurf für die Testspezifikationen vorliegt, wird Mustermaterial erstellt. Dies kann nach den Anweisungen in Kapitel 3 dieses Handbuchs geschehen. Informationen über diese ersten Aufgabenentwürfe können auf vielerlei Art eingeholt werden:

- Durchführung eines **Pilottests** (d. h. einige Personen legen den Test ab) und einer Analyse der Lösungen (siehe Kapitel 3.4 und Anhang VII)
- Befragung von Kolleginnen und Kollegen
- Befragung anderer Beteiligter

Pilottests sollten möglichst mit Personen durchgeführt werden, die der Zielgruppe der Prüfung angehören oder ihr zumindest ähnlich sind. Ein solcher Pilottest sollte zudem unter Prüfungsbedingungen durchgeführt werden, also wie ein Echttest. Die Pilotierung ist aber auch dann noch sinnvoll, wenn die Durchführung unter Echttest-Bedingungen nicht nachgestellt werden kann (vielleicht ist nicht genügend Zeit, um einen vollständigen Test durchzuführen o. Ä.) oder wenn die Anzahl der Teilnehmenden eher klein ist. Auch eine solche Pilotstudie erbringt Hinweise zu den zeitlichen Vorgaben für die einzelnen Aufgaben, über die Klarheit der Anweisungen, über das richtige Layout der Antwortmöglichkeiten etc. Für die mündlichen Prüfungsteile empfiehlt sich die Beobachtung von Teilnehmerleistungen (z. B. mithilfe einer Aufnahme).

Die Befragung von Kollegen und anderen Beteiligten kann ebenfalls auf verschiedene Art und Weise erfolgen. Bei kleinen Gruppen kann man direkt mit den Personen sprechen, wohingegen sich bei größeren Vorhaben Fragebögen oder Feedback-Berichte eignen.

Die Informationen aus der Pilotierung ermöglichen auch das Erstellen umfassender **Lösungsschlüssel** und **Bewertungsskalen** (siehe Kapitel 5.1.3 zu Merkmalen einer Bewertungsanleitung). In den produktiven Teilnehmerleistungen können bestimmte Merkmale identifiziert werden, die der Veranschaulichung der

Kompetenzstufen dienen. Diese Merkmale bilden die Basis der Deskriptoren zur Bewertung für jede Stufe. Auch die Bewertungsskala muss dann pilotiert werden, so dass man das Bewerterverhalten qualitativ oder auch quantitativ analysieren kann (siehe Anhang VII). Weitere Pilotierungen und Überarbeitungen sind ggf. nötig.

Oft sind weitere Untersuchungen erforderlich, um Fragen aus der Pilotierungsphase zu beantworten. Manchmal können die Erkenntnisse aus den Pilottests bereits zur Beantwortung herangezogen werden; in anderen Fällen sind weiterführende Studien durchzuführen. Zum Beispiel:

- Funktionieren die eingesetzten Aufgabentypen bei der speziellen Zielgruppe, für die der Test entwickelt wurde, z. B. Kinder?
- Sind die Aufgabentypen valide in dem Verwendungsbereich, für den der Test konstruiert wurde, z. B. im Tourismus oder im juristischen Bereich?
- Prüfen die Items und Aufgaben tatsächlich die Fertigkeiten, die sie prüfen sollen? Mit statistischen Auswertungen lässt sich bestimmen, wie gut Items und Aufgaben die einzelnen Fertigkeiten prüfen (siehe Anhang VII).
- Können die Bewerter die Bewertungsskalen und die Bewertungskriterien interpretieren und bestimmungsgemäß anwenden?
- Im Falle einer Revision des Testformats: Werden Untersuchungen zur Vergleichbarkeit benötigt, um sicherzustellen, dass das neue Testformat ähnlich funktioniert wie das bestehende?
- Bewirken die Items und Aufgaben bei den Teilnehmenden den beabsichtigten mentalen Prozess? Dies kann mithilfe von Laut-Denken-Protokollen überprüft werden, d. h. die Prüfungsteilnehmerinnen und -teilnehmer äußern ihre Gedankengänge, während sie eine Aufgabe bearbeiten.

Die Testspezifikationen müssen ggf. mehrmals revidiert werden, bevor sie eine für den Echttest verwendbare Form haben.

2.6 Information der Beteiligten

Testspezifikationen erfüllen zahlreiche Zwecke: Sie werden zu Rate gezogen, um Items zu schreiben, um sich auf den Test vorzubereiten und um über Unterrichtsinhalte zu entscheiden. Dies führt in der Regel dazu, dass für die jeweiligen Leser verschiedene Versionen entwickelt werden müssen. So kann zu Vorbereitungszwecken eine vereinfachte Version erstellt werden, in welcher sprachliche Inventare, Themen, Format etc. aufgeführt werden. Eine sehr viel detailliertere Dokumentation muss für Testautorinnen und -autoren erstellt werden.

Zusätzlich zu den Testspezifikationen ist ein Modelltest für die Beteiligten stets von großem Nutzen (siehe Kapitel 3 für mehr Informationen zur Erstellung von Materialien). Wenn notwendig, sollte dieses Mustermaterial nicht nur in Papierform vorliegen, sondern auch Film- und Tonmaterial zum Hörverstehen umfassen. Im Unterricht kann dieses Material zu Vorbereitungszwecken genutzt werden. Später können auch veraltete Tests als Übungstests verwendet werden.

Im Bereich der Aufgaben zum mündlichen und schriftlichen Ausdruck kann es für die Prüfungsteilnehmerinnen und -teilnehmer von Nutzen sein, Beispiellösungen für die Modelltest-Aufgaben vorgelegt zu bekommen. Diese können bei Erprobungsläufen oder aus dem Einsatz älterer Testversionen gewonnen werden. Alternativ können Tipps für Teilnehmende erstellt werden, um ihnen bei der Vorbereitung auf die Prüfung zu helfen.

Das gesamte Material muss verfügbar sein, bevor es tatsächlich benötigt wird. Auch andere Materialien, wie z. B. Prüfungsregularien, Angaben zur Aufgabenverteilung und zu Verantwortungsbereichen sowie Zeitpläne müssen, soweit benötigt, im Vorfeld erstellt werden.

2.7 Schlüsselfragen

- Wer hat entschieden, dass ein Test benötigt wird? Was kann derjenige über den Zweck und Nutzen des Tests sagen?
- Welche Wirkung hat der Test auf die Bildungslandschaft und die Gesellschaft?
- Welche Art und welches Niveau der Sprachleistung sollen bewertet werden?
- Welche Art von Testaufgaben ist hierfür nötig?
- Wie sehen die praktischen Voraussetzungen aus (z. B. Räume, Personal)?
- Wer soll am Entwurf der Testspezifikationen und des Modelltests beteiligt sein (Personen mit Fachwissen, Einfluss, Entscheidungsbefugnissen etc.)?
- Wie werden die inhaltlichen, technischen und organisatorischen Merkmale des Tests in den Testspezifikationen dargestellt?
- Welche Art von Informationen muss den Nutzern z. B. in Form von publizierten Testspezifikationen gegeben werden? Wie werden diese zugänglich gemacht?
- Wie kann der Test erprobt werden?
- Wie können alle Beteiligten am besten über den Test informiert werden?

2.8 Weiterführende Literatur

Beispiele für Teilnehmerleistungen, die sich am GER orientieren und die zur Unterstützung für die Testentwickler und zum besseren Verständnis der GER-Kompetenzstufen dienen, stehen in großer Zahl zur Verfügung. Siehe Europarat (2006a, b; 2005), Eurocentres/Federation of Migros Cooperatives (2004), University of Cambridge ESOL Examininations (2004), CIEP/Eurocentres (2005), Bolton, Glaboniat, Lorenz, Perlmann-Balme und Steiner (2008), Grego Bolli (2008), Europarat und CIEP (2009), CIEP (2009).

Musterformate für Testspezifikationen finden sich bei Bachmann und Palmer (1996: 335–34), Alderson, Clapham und Wall (1995: 14–17) und bei Davidson und Lynch (2002: 20–32).

Es stehen zahlreiche Raster zur Verfügung, die als Leitfaden zur Beschreibung und zum Vergleich von Aufgaben herangezogen werden können, u. a. von ALTE-Mitgliedern (2005a, b; 2007a, b), Figueras, Juijper, Tardieu, Nold und Takala (2005).

3 Generierung von Testversionen

3.1 Der Prozess der Echttesterstellung

Ziel dieser Phase ist es, Material für die tatsächliche Prüfung zu erstellen, das den Testspezifikationen entspricht und rechtzeitig bereitgestellt wird. Die Generierung von Echttests gliedert sich in drei Hauptphasen, die in Abbildung 7 dargestellt werden.

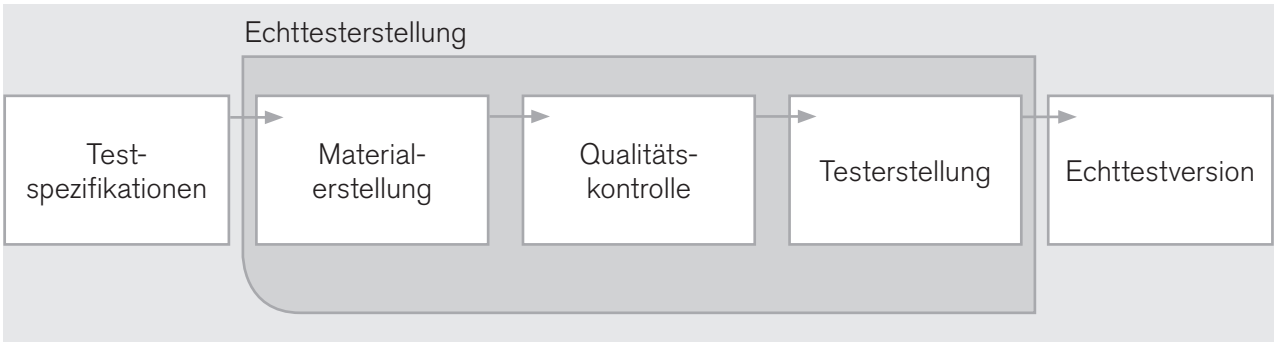


Abb. 7: Hauptphasen des Testerstellungsprozesses

Die Erstellung von Testitems und die Zusammenstellung einer Testversion werden im Folgenden als getrennte Phasen dargestellt, um die Ziele der jeweiligen Schritte besser zu verdeutlichen. Wird Testmaterial produziert und gleich als Testversion zusammengestellt, unterliegen die beiden Schritte jedoch den gleichen Prinzipien. In jedem Fall sollte eine Qualitätskontrolle folgen, die – wenn nötig – noch zu Änderungen am Testmaterial führen kann.

3.2 Erste Schritte

Vor Beginn der Aufgabenerstellung stehen erste Überlegungen:

- Anwerbung und Schulung von Testautorinnen und -autoren
- Verwaltung des Materials

3.2.1 Anwerbung und Schulung von Testautoren

Testautoren und Testentwickler können identisch sein. In diesem Fall ist eine Anwerbung nicht notwendig und die Schulung relativ einfach, da die Autoren bereits mit dem Test und seinen Zielen vertraut sind.

Bei der Suche nach externen Testautorinnen und -autoren muss der Testanbieter entscheiden, welche fachliche Qualifikation diese mindestens mitbringen müssen. Dies betrifft z. B. das Sprachniveau und die Vertrautheit mit dem Prüfungskontext. Weitere wichtige Aspekte der Aufgaben- bzw. Itemerstellung wie Kenntnis über bestehende Tests oder die Prinzipien der Bewertung können im Rahmen der Qualifizierung behandelt werden (siehe ALTE 2005), müssen also nicht vorausgesetzt werden. Regelmäßige Schulung, Kontrolle und Evaluierung tragen zur kontinuierlichen professionellen Entwicklung der Testautorinnen und -autoren bei.

Sprachlehrkräfte sind oft gute Testautoren, da sie ein vertieftes Verständnis für Sprachlernende und für die Sprache selbst entwickelt haben. Sie sind besonders gut geeignet, wenn sie Lernende bereits auf

ähnliche Tests vorbereitet haben oder bei der Bewertung oder bei mündlichen Prüfungen mitwirken. Testautorinnen und -autoren erstellen entweder alle Teile des Tests oder arbeiten nur an speziellen Teilbereichen, je nach Wissensstand und nach Anforderungen des Testanbieters.

3.2.2 Verwaltung des Materials

Der Testanbieter muss ein System entwickeln, wie Testitems eingereicht, gespeichert und bearbeitet werden. Dies ist vor allem bei einer großen Anzahl von Items und Aufgaben erforderlich. Das gesamte Testmaterial muss den gleichen Prozess der Qualitätssicherung durchlaufen, u. a. mit Redaktion und Erprobung. Daher sollte man jederzeit nachvollziehen können, in welcher Prozessphase sich das einzelne Item gerade befindet. Je mehr Items produziert werden und je mehr Personen in den Prozess involviert sind, desto wichtiger wird dieser Aspekt. Ein einfaches System für die Verwaltung des Materials sollte Folgendes enthalten:

- eine eigenständige Identifikationsnummer für jedes Item
- eine Checkliste, die die jeweilige Prozessphase, die Änderungen und andere Informationen zum Item aufzeigt
- ein System, das die Items und die entsprechenden Informationen jederzeit zugänglich macht und garantiert, dass keine überholten Varianten aus der Erstellungsphase in den Echteinsatz gelangen – ggf. durch Ablage an einem gesonderten, zentralen Ort oder durch Weiterleitung aller Informationen per E-Mail am Ende jeder Schreib- und Bearbeitungsphase.

3.3 Itemerstellung

Testautorinnen und -autoren erhalten die Aufgabe, Material zur Verwendung in Echttests zu produzieren. In diesem Handbuch wird dieser Schritt als „Auftragsvergabe“ bezeichnet. Anhang IV gibt Informationen, die für diesen Schritt hilfreich sein können. Testautoren müssen wissen, wie viele Items benötigt werden, welcher Art sie sein sollen und wann sie vorliegen müssen.

Der folgende Abschnitt befasst sich mit der Frage, welches Material benötigt wird und wie die Kommunikation mit den Testautoren erfolgen soll. Die Entscheidung, wann die Items fertiggestellt werden müssen, trifft man, indem man vom Datum des Ersteinsatzes des Echttests herunterrechnet.

3.3.1 Abschätzung des Bedarfs

Um einen Test zu erstellen, benötigen die Testanbieter eine Auswahl von Aufgaben und Items. Wie viele Items in Auftrag gegeben werden sollen, ist schwer zu bestimmen, da jede Testversion verschiedene Aspekte in angemessener Weise zusammenführen muss: Aufgabentypen, Themen, sprachlicher Fokus und Schwierigkeitsniveau (siehe Kapitel 3.5). Daher müssen mehr Items in Auftrag gegeben werden, als tatsächlich verwendet werden. Auch kann man davon ausgehen, dass einige Items bei der Qualitätskontrolle aussortiert werden.

3.3.2 Auftragsvergabe

Materialien können für einen bestimmten Prüfungstermin oder als Ergänzung für eine **Itemdatenbank** in Auftrag gegeben werden, aus der die Testversionen später zusammengestellt werden. In beiden Fällen muss genügend Zeit für die Erstellung zur Verfügung stehen.

Für das Erstellen von Items sollten bestimmte Parameter vereinbart und festgeschrieben werden, damit es nicht zu Irritationen oder Missverständnissen kommt. Eine längere Liste von offiziellen Anforderungen ist sinnvoll, wenn die Gruppe der Testautoren relativ groß und heterogen ist. Die Übernahme einiger Punkte aus folgender Auswahl ist aber in jedem Fall von Nutzen:

Angaben zum benötigten Material

Informationen:

- die benötigte Anzahl an Texten, Aufgaben und Items
- für Texte: ob die dazu gehörigen Items gleich geschrieben werden sollen oder erst nach Abnahme des Texts
- für mündliche Tests mit visuellen Vorgaben: ob Bilder erwartet werden oder nur die Angabe darüber, welche Art Bilder benötigt wird
- Copyright für Bilder und Texte und der Umgang damit

Anforderungen:

- Lösungsschlüssel für jedes Item sowie Bewertungshinweise
- für Schreibaufgaben: Beispiellösungen, um sicherzustellen, dass die Aufgaben mit der vorgegebenen Mindestwortzahl und den Sprachkenntnissen der jeweiligen Teilnehmergruppe bewältigt werden können
- Ausfüllen eines standardisierten Formulars zur Beschreibung der Aufgabe

Angaben zur Aufmachung des Materials

- Am sinnvollsten ist die elektronische Form, denn so kann das Material leicht abgespeichert werden. Bei der Erstellung verwendet man am besten eine Formatvorlage, die für weitere Items ein gleich bleibendes Format sicherstellt.
- Wenn eine vollständige Testversion geschrieben wird, sollte geklärt sein, ob die Items durchgehend nummeriert und die Testteile aufeinander folgen sollen oder ob jeder Testteil oder jede Aufgabe auf einem separaten Blatt stehen soll.
- Ebenfalls muss festgelegt werden, welche Angaben die Testautorinnen und -autoren bei der Einreichung des Materials machen sollen, z. B. Name des Testautors, Datum und Bezeichnung des Tests.

Alle diese Angaben können durch den Leitfaden für Testautoren abgedeckt werden – siehe unten.

Angaben zur Abgabefrist

Die Testautorinnen und -autoren müssen wissen, wann ihr Material redigiert wird und ob sie bei der Redaktionssitzung anwesend sein sollen. Auch wenn die Testautoren nicht in den weiteren Testerstellungsprozess involviert werden, so sollte man ihnen dennoch mitteilen, wie ihre Rolle in den gesamten Produktionsprozess passt, da dies hilfreich ist, um die gesetzten Fristen zu verstehen und einzuhalten.

Weitere Angaben, z. B. Arbeitsvertrag

Arbeiten die Testautorinnen und -autoren auf freiberuflicher Basis oder schreiben sie die Items neben anderen Tätigkeiten für den Testanbieter, so müssen sie über die Bedingungen ihrer Beschäftigung informiert werden. Möglicherweise wird nur für Material bezahlt, das akzeptiert wird (für verworfenes Material dagegen nicht). Alternativ erfolgt eine Abschlagszahlung bei Einreichung des Materials und eine zweite Zahlung nach dessen Annahme. Die Bezahlung ist je nach Itemtyp normalerweise unterschiedlich. Honorarvereinbarungen betreffen eine vollständige Testversion oder Teilabschnitte.

Wenn Lehrerinnen und Lehrer Material für Schulprüfungen schreiben sollen, so muss ihnen neben ihren Unterrichtsverpflichtungen genügend Zeit zur Materialerstellung eingeräumt werden.

Die folgenden Unterlagen sollten den Testautorinnen und -autoren zur Verfügung stehen:

- ausführliche Testspezifikationen für Testautoren. Dies sind ggf. vertrauliche Unterlagen mit detaillierteren Angaben als die öffentlich zugänglich gemachten Informationen und mit genauen Hinweisen zur Auswahl und Aufmachung des Materials. So verhindert man, dass Testautoren Zeit mit eigenen – ggf. falschen – Vermutungen verschwenden, was akzeptabel ist.
- Modelltests oder ältere Testversionen

Testautorinnen und -autoren müssen über die Zielgruppe der Prüfung informiert werden, also u. a. über Alter, Geschlecht und sprachlichen Hintergrund.

Weitere Unterlagen oder Informationen sind je nach Kontext nötig, zum Beispiel:

- ein vom Testautor zu unterschreibendes Auftragsvergabeformular
- ein Vertrag, der dem Testanbieter das Copyright auf die Testmaterialien zusichert
- eine Lexikliste, die die Bandbreite und das Niveau des Vokabulars und/oder der zu verwendenden Strukturen zeigt
- ein Handbuch mit Informationen über den Testanbieter

3.4 Qualitätskontrolle

3.4.1 Redaktion des neuen Materials

Nachdem die Materialien eingereicht wurden, müssen sie auf ihre Qualität hin überprüft werden. Dies geschieht durch Einholen von Expertenmeinungen und durch Erprobung. Wenn ein Item oder eine Aufgabe einmal oder mehrmals geändert wird, muss daraufhin eine weitere Überprüfung stattfinden. Eine solche Überprüfung ist absolut notwendig und sollte idealerweise nicht vom Testautor selbst durchgeführt werden. Bei Mittelknappheit können Items und Aufgaben in einem kleinen Kreis von Kollegen geprüft werden. Sollte der Testautor oder die Testautorin tatsächlich alleine arbeiten, so muss eine zeitliche Pause zwischen dem Schreiben der Items und dem Überprüfen eingeplant werden. Auch trägt es zur Objektivität bei, wenn eine größere Anzahl von Items gleichzeitig überprüft wird.

Zuerst sollte im Rahmen einer **Aufgabenvorrevision** überprüft werden, ob das Material den Testspezifikationen und sonstigen Anforderungen aus der Auftragsvergabe entspricht. Die Testautorinnen und -autoren sollten hierüber eine Rückmeldung erhalten, so dass sie ihre Arbeit überarbeiten und ihre Fertigkeiten in der Itemerstellung weiterentwickeln können. Diese Rückmeldung kann auch Vorschläge zur Änderung eines Items enthalten (siehe Anhang V).

Texte werden auch ohne Items in Auftrag gegeben. Wenn ein Text akzeptiert wurde, schreiben Testautoren in Anschluss daran die Items.

Die erste Überprüfung kann man relativ kurz halten und somit relativ viele Items recht schnell beurteilen. Bei einer großen Anzahl von Items empfiehlt sich hierfür eine besondere Sitzung.

Der zweite Schritt der Itemerstellung erfordert eine genauere Redaktion. Es ist wichtig, dass jedes Item und jede Aufgabe von einer anderen Person als dem Autor selbst kontrolliert wird. Im Schulkontext können z. B. Lehrerinnen und Lehrer gegenseitig ihre Items überprüfen, die sie für ihre eigene Klasse erstellen.

Wenn an einer Redaktionssitzung mehr als vier oder fünf Personen teilnehmen, verlangsamt sich der Prozess in der Regel, wogegen weniger als drei Personen möglicherweise nicht genügend Standpunkte einbringen. Wenn mehrere Sitzungen einberufen werden, sollte ein Vertreter des Testanbieters als Koordinator bestimmt werden, der über den Zeitpunkt, die Auswahl der Teilnehmenden und das zu überprüfende Material bestimmt.

Die Mitglieder der Redaktionsgruppe sollten das Material bereits im Vorfeld sichten, so dass in der Sitzung Zeit gespart wird. Dabei ist auf Folgendes zu achten:

- **Textbasierte Items** sollte man vor dem Lesen des Texts anschauen, so dass man gleich feststellt, welche Items ohne Lesen des Textes beantwortet werden können (z. B. ausschließlich mithilfe von Weltwissen oder anderem Hintergrundwissen).
- Alle anderen Items sollten wie unter Prüfungsbedingungen gelöst werden, ohne auf den Lösungsschlüssel zu schauen. Auf diese Weise können Items identifiziert werden, für die mehr als eine richtige Antwort möglich ist, die unklar oder schlecht formuliert sind, die einen unlogischen **Distraktor** haben, zu schwierig oder nicht verständlich sind.
- Lese- und Hörverstehenstexte sollten auf Länge sowie auf Angemessenheit von Themen, Stil und Sprachniveau hin überprüft werden. Für die Überprüfung des Sprachniveaus wird Expertenwissen benötigt; dieses kann durch das Hinzuziehen linguistischer Beschreibungen unterstützt werden.

Wenn die Redaktion in Gruppen stattfindet, können die erkannten Probleme ausführlich innerhalb der Gruppe angesprochen und diskutiert werden. Oft gibt es rege Diskussionen über das geschriebene Material. Testautorinnen und -autoren müssen in der Lage sein, konstruktive Kritik anzunehmen und zu üben, was sich manchmal als schwierig erweist. Wenn ein Testautor seine Entwürfe erfahrenen Kollegen gegenüber rechtfertigen oder erklären muss, so kann man bereits davon ausgehen, dass es Schwachstellen gibt.

Eine Person innerhalb der Gruppe muss alle Entscheidungen detailliert und genau protokollieren, so dass sämtliche Änderungen der Redaktionsgruppe klar zu erkennen sind. Am Ende der Sitzung darf es keinen Zweifel darüber geben, auf welche Änderungen man sich geeinigt hat.

Die endgültigen Entscheidungen unterliegen dem Testanbieter, der auch bestimmen sollte, wann eine Diskussion beendet ist.

Folgende Punkte sind in einer Redaktionssitzung zu beachten:

- Besonderes Augenmerk sollte auf die **Anweisungen**, die den Teilnehmenden zusammen mit den Items gegeben werden, und die Lösungsschlüssel gerichtet werden.
- Items, die zu einer Verzerrung der Testergebnisse (*Bias*) führen können, lassen sich mithilfe von Listen aller Themen oder anderer Aspekte identifizieren, die es zu vermeiden gilt (siehe Anhang VII).
- Einige Vorschläge haben möglicherweise ein gewisses Potenzial, erfordern aber mehr Änderungen, als in der Redaktionssitzung möglich sind. Diese werden dem Testautor zur Überarbeitung zurückgegeben oder an jemanden mit mehr Erfahrung zur Revision überreicht.
- Nach der Sitzung sollten übrig gebliebene sowie während der Sitzung verwendete Unterlagen aus Sicherheitsgründen vernichtet werden. Die geänderten und akzeptierten Materialien verbleiben beim Testanbieter.
- Testautorinnen und -autoren sollten vom Testanbieter eine Rückmeldung über nicht akzeptiertes Material erhalten, vor allem, wenn sie noch nicht an einer Redaktionssitzung teilgenommen haben oder bei der Sitzung zur Redaktion ihres eigenen Materials nicht anwesend waren.

- Redaktionssitzungen sind eine hervorragende Gelegenheit für neue Testautoren, von erfahrenen Autoren innerhalb der Gruppe mehr über ihre Arbeit zu lernen.

3.4.2 Pilotierung, Vorerprobung und Erprobung

Testitems müssen in irgendeiner Form erprobt werden, u. a. weil Teilnehmende immer wieder unerwartete Lösungen für Testaufgaben finden. In dieser Phase werden je nach Ziel und Mitteln des Testanbieters ein **Pilottest**, eine **Vorerprobung** und eine **Erprobung** oder eine Kombination aus diesen Methoden organisiert.

Im Pilottest wird eine kleine Personengruppe gebeten, die Aufgaben wie in einer tatsächlichen Prüfung zu lösen. Dies kann ganz informell geschehen, z. B. im Kollegenkreis, wenn sich niemand anderes findet. Die **Lösungen** werden analysiert und Kommentare gesammelt (siehe Anhang VI), um das Item weiter zu verbessern.

Erprobungen werden normalerweise für Tests mit **geschlossenen Aufgaben** durchgeführt. Bei der Erprobung werden die Bedingungen der tatsächlichen Prüfung eingehalten. Erprobungsteilnehmerinnen und -teilnehmer werden entsprechend der Zielgruppe ausgewählt. Am Ende müssen genügend gelöste Aufgaben vorliegen, um eine statistische Analyse erstellen zu können (siehe Anhang VII). Diese Analyse zeigt, wie gut die einzelnen Optionen funktionieren, wie schwierig ein Item war, was das durchschnittliche Ergebnis ist, ob der Test das richtige Sprachniveau für die Zielgruppe hat, wie hoch der Fehleranteil war, ob Items zu Verzerrungen des Ergebnisses führen (siehe Anhang VII), ob sie das gleiche Konstrukt messen, wie schwer der Test insgesamt ist etc. Schon die einfachsten statistischen Analysen (siehe Anhang VII) können äußerst informativ sein und mithilfe kostengünstiger und benutzerfreundlicher Software durchgeführt werden.

Lösungen für **offene Aufgaben** (zum Schreiben und Sprechen) können ebenfalls statistisch analysiert werden. Allerdings liefert die qualitative Analyse auch weniger Teilnehmer-Lösungen oft mehr Informationen. Eine Erprobung produktiver Aufgaben in kleinerem Umfang wird auch als Vorerprobung oder *Trialling* bezeichnet, um eine Abgrenzung zum Erproben geschlossener Aufgaben zu finden. Eine solche Methode soll zeigen, ob eine Testaufgabe zufriedenstellend funktioniert und die beabsichtige Leistung elizitiert.

Anders als beim Pilottest läuft die Erprobung wie eine Echtprüfung ab, so dass bestimmte Voraussetzungen erfüllt sein müssen:

- eine ausreichende Zahl an Teilnehmenden (siehe Anhang VII)
- sicher hergestellte Aufgabenhefte
- einen Prüfungsraum und Prüfungspersonal
- Personal zur Auswertung

Die an der Erprobung Teilnehmenden sollten der Prüfungszielgruppe so ähnlich wie möglich sein. Deshalb bietet es sich an, Lernende anzusprechen, die sich gerade auf die entsprechende Prüfung vorbereiten.

Eine Rückmeldung zu den erbrachten Leistungen ist eine gute Möglichkeit, um die Erprobungsteilnehmerinnen und -teilnehmer zu motivieren und ihre tatsächlichen Fähigkeiten abzurufen. Ein Feedback ermöglicht den Lernenden und den Lehrkräften, den aktuellen Stand der Sprachkompetenz festzustellen und die Bereiche zu erkennen, die vor Teilnahme an der Prüfung noch verbessert werden müssen.

Ein möglicher Nachteil beim Einsatz solcher Teilnehmenden besteht allerdings in der Bekanntmachung von Items, was die Prüfungssicherheit im späteren Echteinsatz gefährden kann. Dieses Problem gestaltet bei einigen Testanbietern die Durchführung von Erprobungen sehr schwierig.

Um das Risiko zu minimieren, müssen die Erprobungsaufgaben nicht unbedingt in der gleichen Form präsentiert werden wie der Test selbst. Der Zeitrahmen für die Itemerstellung sollte eine ausreichend lange Zeitspanne zwischen der Erprobung und dem Echteinsatz vorsehen. Wenn Erprobungen außerhalb der Institution des Testanbieters durchgeführt werden, muss das Personal genaue Anweisungen zur sicheren Verwahrung des Materials erhalten und eine Vertraulichkeitserklärung unterschreiben.

Ein Aufgabenheft für die Erprobung muss jedoch der Zusammenstellung im tatsächlichen Test nicht ähneln, da die Items getestet werden sollen und nicht das Testformat. Es steigert allerdings die Motivation der Teilnehmenden bei der Erprobung, wenn sie wissen, dass die Erprobung der tatsächlichen Prüfung sehr ähnlich ist, dies also eine gute Art der Vorbereitung darstellt. Daher ist es ratsam, zumindest ein ähnliches Format wie in der echten Prüfung zu verwenden.

In jedem Fall muss die Erprobung wie der tatsächliche Test ablaufen. Wenn Teilnehmende abgelenkt sind, täuschen oder andere Zeitvorgaben erhalten, führt dies zu schwer auswertbaren Daten.

Ist die Qualität der statistischen Informationen von besonders hoher Bedeutung (z. B. wenn die Items **kalibriert** werden – siehe Anhang VII), muss die Anzahl der Erprobungsteilnehmerinnen und -teilnehmer entsprechend groß sein. Die anzustrebende Mindestanzahl hängt von der Art der geplanten Analyseverfahren ab, aber selbst eine kleine Gruppe von Probanden (unter 50 Personen) kann zu relevanten Informationen führen und auf Schwierigkeiten bei Items hinweisen. Bei kleineren Gruppen bietet sich eher eine qualitative Analyse an.

Wichtiger noch als die Gruppengröße ist es, Teilnehmende zu finden, die denjenigen in der Echtprüfung so ähnlich wie möglich sind. Kleinere, weniger repräsentative Gruppen von Probanden erlauben lediglich vorsichtige Schlussfolgerungen, und die Analyse muss durch die bei der Überprüfung der Items entstandenen Expertenmeinungen ergänzt werden. Siehe Anhang VII für mehr Informationen zu Analysen.

Wird die Erprobung durchgeführt, um qualitative Informationen über die Items zu erhalten, müssen einige Überlegungen angestellt werden, um ein Maximum an nützlichen Anhaltspunkten zu sammeln:

- Für geschlossene Items kann ein Kommentar der Teilnehmenden und Lehrkräfte eingeholt werden. Zur Vereinfachung sollte hier eine Liste von Fragen oder ein Fragebogen verwendet werden (siehe Anhang VII).
- Für Aufgaben in der mündlichen Prüfung, die einen Prüfer als Gesprächspartner (Interlokutor) erfordern, kann auch dessen Rückmeldung sinnvoll sein. Sie zeigt dem Testanbieter, ob die Aufgabe von den Lernenden verstanden wurde, ob sie für ihren Lebensbereich und ihre Altersgruppe angemessen war und ob sie ausreichend Informationen bot, um angemessen ausgeführt werden zu können (siehe Anhang VI).
- Bei offenen Items und Aufgaben sieht man anhand der Teilnehmerleistungen, ob Diskursmittel, syntaktische Strukturen und Wortschatz tatsächlich gezeigt werden konnten, wie für das angezielte Sprachniveau zu erwarten ist.
- Auch Kommentare der Teilnehmenden zur Erprobung insgesamt sowie zu damit in Verbindung stehenden Themen kann man einholen (siehe Anhang VI).

3.4.3 Überprüfung der Items

Auf den Pilottest oder die Erprobung sollte eine erneute redaktionelle Bearbeitung der Items folgen. An einer solchen Sitzung nehmen gewöhnlich die Testanbieter und erfahrene Testautoren teil, bei der Besprechung offener Items und Aufgaben (z. B. Schreibaufgaben) auch erfahrene Bewerter.

Ziel dieser Sitzung ist es, aufgrund der Ergebnisse des Pilottests bzw. der Erprobung zu entscheiden, welche Items beibehalten werden, welche verändert und welche abgelehnt werden. Dies wird inklusive einer Nachbesserung und erneuten Pilotierung oder Erprobung in Abbildung 8 dargestellt.

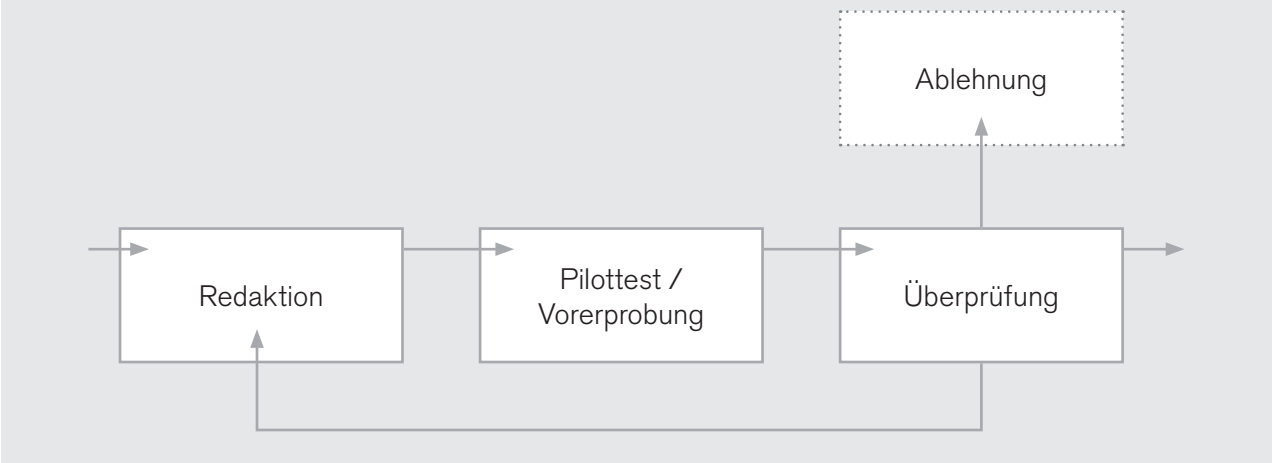


Abb. 8: Itemverbesserung durch Qualitätskontrolle

Überprüfungen nach der Erprobung sollten die folgenden Punkte abdecken:

- Welches Material kann für Echttests verwendet werden?
- Welches Material muss als ungeeignet abgelehnt werden?
- Welches Material muss umgeschrieben und noch einmal getestet werden, bevor es für die Aufnahme in den Echtbetrieb in Betracht gezogen wird?

In der Besprechung sollte auf Folgendes geachtet werden:

- Inwieweit entsprachen die Erprobungsteilnehmerinnen und -teilnehmer der Zielgruppe des tatsächlichen Tests? Dies gibt eine Vorstellung darüber, wie vertrauenswürdig Analyseergebnisse sind.
- Fanden die Teilnehmerinnen und Teilnehmer einen Zugang zu den Aufgaben und Themen? Gab es organisatorische Probleme?
- Die Güte der einzelnen Items und Aufgaben. Wenn offene Aufgaben evaluiert werden, ist es hilfreich, eine Auswahl verschiedener Teilnehmerleistungen zur Verfügung zu haben. Bei geschlossenen Items zeigt eine statistische Analyse ggf. Probleme mit Items auf, die bei einer Überprüfung durch Experten bestätigt und korrigiert werden können. Basiert die Analyse jedoch auf unzureichenden Daten (z. B. ungeeignete oder nur wenige Teilnehmende) muss man Vorsicht walten lassen. In diesem Fall sollte anderen Informationen, wie etwa einer qualitativen Einschätzung der Items und Aufgaben, mehr Bedeutung zugemessen werden.
- Wenn statistische oder andere Informationen auf Aufgaben aus einer Itembank angewendet werden, muss eine einheitliche und konsequente Vorgehensweise eingehalten werden. Nur so entsteht ein nützliches Werkzeug für die Generierung von Testversionen. Anhang VII gibt weitere Informationen zur statistischen Analyse.

3.5 Erstellung von Testversionen

Wenn ausreichend Material zur Verfügung steht, können Testversionen zusammengestellt werden.

In der Phase der Versionserstellung müssen mehrere Aspekte gegeneinander abgewogen werden – etwa Testinhalt und Itemschwierigkeit –, so dass die jeweilige Testversion als Ganzes den Anforderungen der Testspezifikationen gerecht wird.

Bestimmte Merkmale der Testversion können auf Grundlage der Testspezifikationen und des Formats festgelegt werden (z. B. Anzahl und Typ der Items und Aufgaben), während andere Merkmale höhere Flexibilität innerhalb bestimmter Grenzen verlangen (z. B. Themen, verschiedene Akzente etc.). Ein Leitfaden hilft dabei, eine angemessene Ausgewogenheit zwischen den folgenden Merkmalen sicherzustellen:

- Schwierigkeitsstufe (Diese kann subjektiv beurteilt werden. Wenn jedoch Itembanken genutzt werden, kann man auch mit einem **Mittelwert** und einem Spektrum der Schwierigkeit für die Testitems arbeiten, siehe Anhang VII.)
- Inhalt (Themen oder Handlungsfelder)
- Abstufung (wenn der Test nach und nach schwieriger wird)

Diesen Leitfaden muss man auf die gesamte Testversion anwenden, indem man sie einer Gesamtbetrachtung unterzieht und die Testteile miteinander vergleicht.

Für einige Testformate sind weitere Überlegungen notwendig. Wenn beispielsweise ein Test zum Leseverstehen mehrere Texte enthält, muss man darauf achten, dass sich die Themen nicht wiederholen und dass die Gesamtlänge die vorgegebene Anzahl der Wörter nicht überschreitet. Bei einem Test zum Hörverstehen muss wiederum auf eine Ausgewogenheit zwischen männlichen und weiblichen Stimmen und auf regionale Akzente (soweit zutreffend) geachtet werden.

3.6 Schlüsselfragen

- Wie wird der Prozess der Itemerstellung organisiert?
- Kann eine Itemdatenbank benutzt werden?
- Wer schreibt das Material?
- Was sind die Anforderungen an die Qualifikation der Testautorinnen und -autoren?
- Welches Training erhalten sie?
- Wer nimmt an den Redaktionssitzungen teil?
- Wie werden Redaktionssitzungen organisiert?
- Kann Material erprobt werden?
- Welche Konsequenzen kann es geben, wenn Material nicht erprobt wird, und wie wird mit diesen umgegangen?
- Welche Analyse wird mit den Daten aus der Erprobung durchgeführt?
- Wie wird die Analyse verwendet (z. B. für die Zusammenstellung einer Testversion, für Schulungen von Testautoren etc.)?
- Wer ist an der Generierung einer Testversion beteiligt?

- Welche Variablen müssen berücksichtigt und miteinander in Einklang gebracht werden (z. B. Schwierigkeitsgrad, thematischer Inhalt, Bandbreite der Aufgabentypen etc.)?
- Welche Rolle spielen statistische Analysen (z. B. zur Ermittlung der mittleren Schwierigkeit und des Schwierigkeitsspektrums)?
- Wie wichtig sind diese statistischen Analysen in Relation zu anderen Informationen, wenn Entscheidungen getroffen werden müssen?
- Wird die neu erstellte Testversion noch einmal unabhängig überprüft?
- Wie wird die neue Testversion mit formatgleichen anderen Testversionen abgeglichen bzw. in eine größere Menge von Testversionen eingegliedert?

3.7 Weiterführende Literatur

Leitfaden für Testautoren, siehe ALTE (2005)

Analyse von Testaufgaben, siehe ALTE (2004a, b, c, d, e, f, g, h, i, j, k).

Linguistische Beschreibungen einiger Sprachen, die auf den GER Bezug nehmen, sind für einige Sprachen verfügbar: *Reference Level Descriptors* (RLDs) (Beacco und Porquier (2007, 2008), Beacco, Bouquet und Porquier (2004), Glaboniat, Müller, Rusch, Schmitz und Wertenschlag (2005), Instituto Cervantes (2007), Spinelli und Parizzi (2020), www.englishprofile.org). *Threshold* (van Ek und Trim 1991), *Waystage* (van Ek und Trim 1990) und *Vantage* (von Ek und Trim 2001) sind Vorläufer der RLD.

Anhang VII bietet weitere Informationen über die Möglichkeiten, statistische Daten für die Erstellung einer Testversion zu nutzen.

4 Prüfungsdurchführung

4.1 Ziele der Prüfungsdurchführung

Ziel des Einsatzes einer Testversion im Rahmen einer Prüfungssituation ist es, genaue und reliable Informationen über die Kompetenz jedes Teilnehmenden einzuholen.

Die große Herausforderung liegt hierbei nicht in der Verbesserung der Qualität des Testmaterials wie in den vorangegangenen Arbeitsschritten, sondern in der Logistik. Testanbieter müssen sicherstellen,

- dass die Leistungen der Teilnehmerinnen und Teilnehmer in der Prüfung soweit wie möglich nur von ihrem Sprachniveau abhängen und nicht durch irrelevante Faktoren, wie z. B. eine laute Umgebung oder Täuschungsversuche, beeinflusst werden,
- dass die Teilnehmer-Lösungen und erste Bewertungen effizient und sicher eingeholt und für die nächste Phase der Auswertung und Bewertung verfügbar gemacht werden können,
- dass das gesamte Testmaterial zur richtigen Zeit am richtigen Ort ist.

Diese Punkte sind sowohl für kleine als auch für umfangreiche Prüfungen wichtig. Manchmal ist etwas Einfaches von entscheidender Bedeutung, zum Beispiel die Eignung des Raums zu überprüfen.

Ein weiteres Ziel kann es sein, mehr Informationen über den Hintergrund der Teilnehmenden einzuholen. Dies ist vor allem wichtig, wenn der Testanbieter die Gruppe der Teilnehmenden noch nicht kennt. Die entsprechenden Informationen führen zu einem besseren Verständnis darüber, wer den Test absolviert, und unterstützen somit den Nachweis über die Validität des Tests (siehe Anhang I und Anhang VII).

4.2 Der Prozess der Prüfungsdurchführung

Der Prozess der Prüfungsdurchführung wird in Abbildung 9 dargestellt. Einige Phasen, wie die Anmeldung der Teilnehmenden oder die Materialauslieferung, sind in einigen Kontexten recht unkompliziert, z. B. bei einem schulinternen Test. Auch hier ist jedoch darauf zu achten, dass die Prüfungsräume angemessen sind und externe Einflüsse, wie z. B. Lärm, vermieden werden. In anderen Kontexten ist die Logistik weitaus schwieriger und verlangt ein höheres Maß an Aufmerksamkeit.

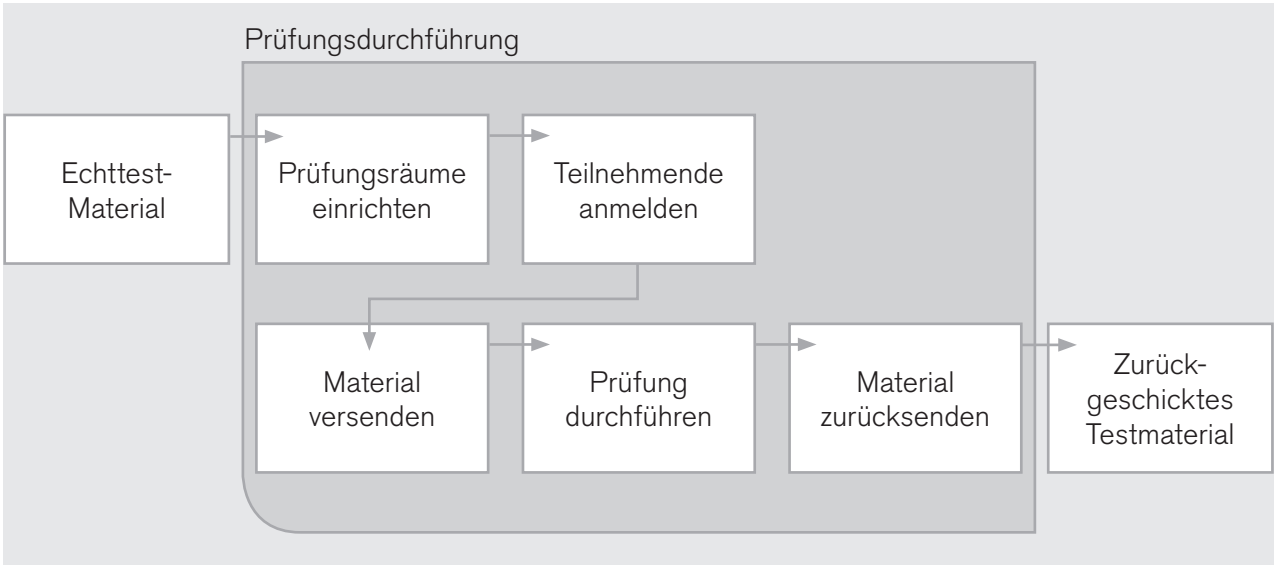


Abb. 9: Prozess der Prüfungsdurchführung

4.2.1 Organisation des Prüfungsortes

Die Räumlichkeiten, in denen die Prüfung durchgeführt wird, sollten im Vorfeld besichtigt werden, entweder durch den Prüfungsanbieter oder einen Dritten, z. B. eine vertrauenswürdige Person am Prüfungsort. Wenn eine andere Institution die Prüfung durchführt, muss sie entsprechend lizenziert sein. Dabei müssen Kriterien wie die folgenden eingehalten werden:

- Kapazität, um die Prüfung mit der erwarteten Anzahl von Teilnehmenden durchzuführen
- Zugang zu den Prüfungsräumen
- Sicherheit der Aufbewahrungsmöglichkeiten
- Bereitschaft, sich an die Regeln des Testanbieters zu halten
- Bereitschaft, Personal für die Einhaltung der Regeln zu schulen

Wenn Prüfungszentren von Dritten verwaltet werden, sollte der Testanbieter einen Kontrollprozess einführen, bei dem stichprobenartig die Qualität der Testdurchführung in seinem Namen kontrolliert wird.

Bei der Überprüfung der Räumlichkeiten greifen immer die gleichen Kriterien. Es ist durchaus sinnvoll, die Räumlichkeiten vor jeder Testdurchführung zu überprüfen, da Veränderungen eingetreten sein können, über die der Testanbieter nicht informiert wurde, z. B. Bauarbeiten in der unmittelbaren Umgebung.

Folgende Punkte sind zu überprüfen:

- Außengeräusche
- Akustik im Inneren des Raums (vor allem für Hörverstehenstests)
- Größe (Möglichkeit, die vorgesehene Anzahl vor Teilnehmenden mit genügend Sitzabstand zu verteilen)
- Raumaufteilung (damit die Aufsichtspersonen alle Teilnehmer gut im Blick haben)
- Zugänglichkeit
- räumliche Ausstattung in Bezug auf Toiletten oder Warteräume für Prüfungsteilnehmerinnen und -teilnehmer u. Ä.
- sichere Aufbewahrungsmöglichkeiten für Prüfungsmaterial vor und nach der Prüfung

Wenn eine Einrichtung sich als grundsätzlich ungeeignet für die Prüfungsdurchführung herausstellt oder gravierende Fehler bei der Durchführung macht, sollte sie von der Liste der möglichen Prüfungsorte oder Prüfungspartner gestrichen werden.

4.2.2 Anmeldung der Teilnehmenden

Findet der Test im Unterrichtskontext statt, so sind die Teilnehmenden bekannt und eine Kurs- oder Klassenliste reicht für die Testdurchführung aus. Wenn die Prüfungsteilnehmerinnen und teilnehmer dem Testanbieter aber unbekannt sind oder wenn sich Externe für die Prüfung anmelden können, müssen Informationen über die Teilnehmenden eingeholt werden. Ein Anmeldeprozess liefert die für die Prüfungsdurchführung sowie für die Verarbeitung und Bekanntgabe der Ergebnisse benötigten Informationen. An diesem Punkt können zudem Teilnehmende mit Behinderungen besondere Prüfungsbedingungen beantragen, z. B. bei

- Gehörlosigkeit oder Hörbeeinträchtigung
- Blindheit oder Sehbehinderung

- Lese-/Rechtschreibschwäche
- motorischer Beeinträchtigung

Anträge auf barrierefreie Prüfungsbedingungen sind sorgfältig zu prüfen und sollten in berechtigten Fällen zu Hilfestellung und/oder einem Ausgleich bei der Bewertung führen. Daher ist es ratsam, ein Standardverfahren für die häufigsten Anforderungen zu definieren. Ein solches Verfahren regelt die Art der zu erbringenden Nachweise (z. B. ein ärztliches Attest), die möglichen Sonderregelungen und die Frist für die Beantragung.

In einigen Fällen ist Barrierefreiheit leicht herzustellen, etwa bei Gehbehinderungen, die eine Hilfestellung beim Einnehmen des Platzes im Prüfungsraum erforderlich machen.

Andere Maßnahmen müssen eingehender abgestimmt werden. Bei Schwierigkeiten im Lesen, sei es durch eine Leseschwäche oder durch eine Sehbehinderung, können etwa besondere Testbögen oder andere Hilfen bereitgestellt werden. Diese Art der Prüfungsdurchführung darf jedoch in Bezug auf die Prüfungsergebnisse niemandem einen Vorteil gegenüber anderen Teilnehmenden erbringen.

In dieser Anmeldephase ist es weiterhin möglich, Hintergrundinformationen über die Prüfungsteilnehmerinnen und -teilnehmer einzuholen. Die Angaben zu den Merkmalen der Teilnehmenden können genutzt werden, um wichtige Rückschlüsse über die Vergleichbarkeit der verschiedenen Teilnehmer-Gruppen zu ziehen. Es handelt sich um Angaben wie die folgenden:

- Bildungshintergrund
- Erstsprache
- Geschlecht
- Alter
- Lernbiographie in der Zielsprache

In jedem Fall müssen die Teilnehmenden darüber informiert werden, warum diese Daten erhoben werden. Alle Daten sind nach Maßgabe des Datenschutzes zu bearbeiten und zu speichern.

Zusätzlich zum Einholen von Informationen dient der Anmeldeprozess auch dazu, die Teilnehmenden zu informieren. Sie müssen die Teilnahmebedingungen zur Kenntnis nehmen, über die Regeln der Prüfungsdurchführung, über Einspruchsmöglichkeiten, barrierefreie Prüfungsbedingungen usw. informiert werden. Auch müssen ihnen Zeit und Ort der Prüfung und alle anderen praktischen Hinweise, die zur Zeit der Anmeldung zur Verfügung stehen, gegeben werden. Diese Informationen sollten in gedruckter Form vorliegen, im Internet abrufbar sein oder als standardisierte E-Mail verschickt werden, so dass alle Teilnehmenden vollständige und korrekte Informationen erhalten.

Die Anmeldung kann durch den Prüfungsanbieter selbst oder vom Prüfungszentrum oder Dritten, wie z. B. dem Bildungsministerium, durchgeführt werden. Der Testanbieter sollte soweit wie möglich sicherstellen, dass alle Prüfungsteilnehmerinnen und -teilnehmer das gleiche Anmeldeverfahren durchlaufen.

4.2.3 Materialversand

Das Material muss möglicherweise zum Prüfungsort geschickt werden. Der Versand muss sowohl rechtzeitig erfolgen als auch sicher sein, so dass alle Materialien am Tag der Prüfung vor Ort sind.

Oft ist es besser, das Material frühzeitig zu versenden. So stellt man sicher, dass es am Tag der Prüfung zur Verfügung steht und dass fehlendes Material ersetzt werden kann. Hierzu muss der Testanbieter aber sicher sein können, dass das Material bis zum Einsatz immer unter Verschluss gehalten wird.

Die Organisatoren müssen den Inhalt der Sendung bei Erhalt prüfen und mit einer Checkliste abgleichen. Wenn etwas fehlt oder beschädigt ist, müssen sie Ersatz- oder zusätzliches Material gemäß einem Standardverfahren nachfordern.

4.2.4 Prüfungstermin

Für den Prüfungstag müssen genügend Aufsichtspersonen, Bewerter und sonstiges Personal zur Verfügung stehen. Jeder, der an der Durchführung der Prüfung beteiligt ist, muss im Vorfeld seine Aufgaben kennen und verstanden haben. Wenn mehrere Räume oder Prüfungsdurchgänge mit entsprechendem Personaleinsatz geplant sind, ist ein klarer Einsatzplan vonnöten.

Die Durchführungsrichtlinien sollten Anweisungen dazu enthalten, wie man die Identität der Prüfungsteilnehmerinnen und teilnehmer überprüft und wie man damit umgeht, wenn jemand zu spät kommt.

Vor Beginn der eigentlichen Prüfung müssen die Teilnehmenden genau darüber informiert werden, wie sie sich während der Prüfung zu verhalten haben. Dabei geht es etwa um nicht genehmigtes Material, Mobiltelefone, das Verlassen des Raums während der Prüfung sowie den Zeitplan. Außerdem muss man auf Konsequenzen im Falle von Regelverstößen hinweisen, z. B. Sprechen und Abschreiben.

Die Aufsichtspersonen müssen wissen, wie sie sich verhalten sollen, wenn während der Prüfung Regeln gebrochen werden oder andere unvorhergesehene Ereignisse auftreten, z. B. wenn Teilnehmende bei Täuschungsversuchen ertappt werden, ein Stromausfall auftritt oder irgendein anderer Vorfall zu Verzerrungen, Unfairness oder Prüfungsabbruch führt. Aufsichtspersonen müssen auch mit Täuschungsmöglichkeiten vertraut sein, die z. B. durch den Gebrauch von digitalen Aufnahmegeräten, MP3-Playern, Scan-Stiften oder Mobiltelefonen mit Kamera entstehen können.

Im Falle unvorhergesehener Ereignisse lautet die Anweisung an die Aufsichtspersonen bzw. den Prüfungsverantwortlichen, sinnvoll nach eigener Urteilskraft zu handeln und dem Testanbieter einen ausführlichen Bericht zu liefern. Dieser Bericht sollte Angaben über die Anzahl der Betroffenen sowie den Zeitpunkt des Vorfalls enthalten und den Vorfall beschreiben. Auch sollte dem Aufsichtspersonal eine Telefonnummer zur Verfügung gestellt werden, die im Notfall für Rückfragen angerufen werden kann.

4.2.5 Rücksendung des Materials

Prüfungsunterlagen müssen verpackt und an den Prüfungsanbieter zurückgeschickt oder vernichtet werden. Im Falle der Rücksendung können auch andere Unterlagen, wie z. B. Anwesenheitslisten oder der Sitzplan, sofort nach der Prüfung beigelegt werden. Das Material muss auf sicherem Wege zurückgeschickt werden, in der Regel so, wie es auch versandt wurde. Der Frachtführer sollte die Möglichkeit einer Nachverfolgung für den Fall von verspäteter oder verloren gegangener Lieferung anbieten.

4.3 Schlüsselfragen

- Welche Ressourcen stehen für die Durchführung der Prüfung zur Verfügung? (Verwaltungspersonal, Aufsichtspersonen, Räumlichkeiten, CD-Abspielgerät etc.)
- Wie soll das Personal geschult werden?
- Wie können Räumlichkeiten, CD-Abspielgerät und dergleichen vor dem Prüfungstag überprüft werden?
- Wie oft finden die Prüfungen statt?

- Wie viele Teilnehmende werden erwartet?
- Wie werden die Prüfungsteilnehmerinnen und -teilnehmer angemeldet und wie wird ihre Anwesenheit kontrolliert?
- Wie viele Prüfungsräume werden genutzt? Wenn es mehrere Prüfungsräume gibt, liegen diese weit auseinander oder sind sie schlecht zu erreichen?
- Wie kommt das Material zu den Prüfungsorten und wieder zurück?
- Wie wird das Material bis zum Prüfungstermin sicher verwahrt?
- Wo können Probleme auftreten? Gibt es Verfahren oder Regeln für einen solchen Fall?

4.4 Weiterführende Literatur

Siehe ALTE (2006b) für eine Checkliste zur Selbstbeurteilung von Logistik und Durchführung.

5 Auswertung, Benotung und Übermittlung der Ergebnisse

Ziel der **Auswertung** ist es, die Leistung jedes Teilnehmenden einzuschätzen und eine genaue und reliable Bewertung für jeden zu erlangen. **Benotung** bedeutet, dass jeder Teilnehmende einer aussagekräftigen Kategorie zugeordnet wird, so dass sein jeweiliges Prüfungsergebnis leicht zu verstehen ist. Eine aussagekräftige Kategorie kann eine der gemeinsamen Referenzstufen des GER sein, wie A2 oder C1. Bei der Übermittlung der Ergebnisse geht es darum, den Prüfungsteilnehmerinnen und -teilnehmern und anderen Beteiligten neben dem Ergebnis der Prüfung auch alle notwendigen Informationen zur angemessenen Interpretation des Ergebnisses mitzuteilen, wenn z. B. eine Entscheidung getroffen werden muss, ob dem Teilnehmenden ein bestimmter Arbeitsplatz angeboten wird. Abbildung 10 zeigt den gewöhnlichen Ablauf des Prozesses. In einigen Fällen kann die Bewertung der Leistung auch zur gleichen Zeit wie der Test selbst stattfinden. Dies ist z. B. manchmal bei mündlichen Prüfungen der Fall, auch wenn hier vor der Bekanntgabe des Ergebnisses die Bewertungen vom Testanbieter ggf. noch angepasst werden.

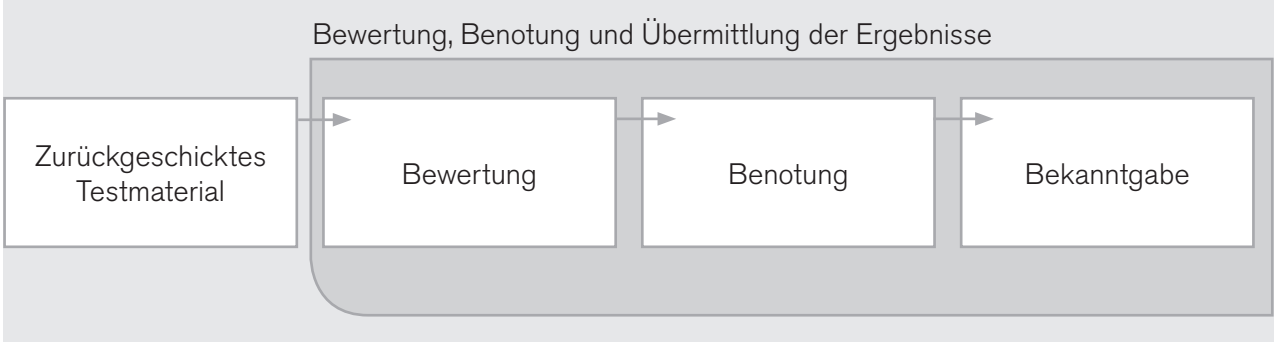


Abb. 10: Der Prozess der Auswertung, Bewertung und Übermittlung der Ergebnisse

Erste Schritte

Vor der Auswertung und der Bewertung müssen die folgenden Schritte unternommen werden:

- die Bewertungsmethode entwickeln
- **Auswerter** und **Bewerter** anwerben
- Auswerter und Bewerter schulen

5.1 Auswertung

Der Begriff „**Auswertung**“ umfasst alle Aktivitäten, die zur Punktevergabe für die Teilnehmer-Lösungen führen. Dabei wird oft unterschieden zwischen dem „**Auswerter**“, der für seine Aufgabe weniger Fachwissen benötigt, und dem „**Bewerter**“, der eine Fachausbildung durchlaufen haben muss. Diese Unterscheidung wird auch hier im Text getroffen. Der folgende Abschnitt behandelt manuelle und maschinelle Auswertung.

5.1.1 Manuelle Auswertung

Personen, die **manuell auswerten**, müssen keine Testexperten sein – hohe Sprachkompetenz in der getesteten Sprache ist ausreichend. Sie benötigen jedoch eine Schulung, Unterstützung und einen eindeutigen Lösungsschlüssel, um ihrer Aufgabe gerecht zu werden.

Der Auswertungsprozess muss so organisiert sein, dass das Verfahren planmäßig verläuft und die Ergebnisse zum gewünschten Zeitpunkt vorliegen. Auch darf die Arbeitsbelastung der Auswerter nicht so hoch sein, dass Reliabilität oder Genauigkeit gefährdet werden.

Auswerter anwerben und schulen

Manuelles Auswerten im einfachsten Sinne bedeutet, die Teilnehmer-Lösungen mit einer oder mehreren vorgegebenen Lösungen abzugleichen. Aufgabentypen wie *Multiple-Choice*-Aufgaben sind das beste Beispiel für den Fall, in dem keine Abweichung von den gegebenen Optionen erlaubt ist. Für die Auswertung solcher geschlossener Aufgaben muss der Auswerter in der Lage sein, die entsprechende Sprache zu lesen, aufmerksam zu arbeiten und gut mit eintöniger Arbeit umzugehen; ansonsten braucht er keine weiteren Fähigkeiten. Schulungen sollten darauf abzielen, die Auswerter mit den Verfahrensweisen vertraut zu machen. Mit entsprechender technischer Ausstattung kann diese Art der Auswertung mindestens genauso gut oder besser maschinell erledigt werden.

Wenn Items mehr als nur den Abgleich mit einer vorgegebenen Lösung erfordern, brauchen die Auswerter mehr Wissen über die Sprache, die Sprachlernenden und die Teststruktur. **Items mit variabler Punkteskala** werden beispielsweise je nach Lösungserfolg unterschiedlich bewertet. So kann z. B. ein Punkt für die Wahl des richtigen Verbs und ein weiterer Punkt für die richtige Verbform vergeben werden. Hier wird mehr Fachwissen von einem Auswerter erwartet als bei der simplen Unterscheidung, ob eine Antwort richtig oder falsch ist.

Für diese Items ist es oft schwierig, einen Lösungsschlüssel bereitzustellen, der alle möglichen Antworten aufzeigt. Daher ist es hilfreich, wenn der Auswerter auch alternative Antworten erkennt und meldet.

Bei nicht fest angestellten, jedoch regelmäßig eingesetzten Auswerterinnen und Auswertern bietet sich ein Beurteilungssystem an, das auf Parametern wie Genauigkeit, Reliabilität und Geschwindigkeit basiert. Auswerter, die nicht zur Zufriedenheit gearbeitet haben, werden ggf. nicht weiter beschäftigt oder neu geschult. Eine solche Schulung kann Teil eines Systems sein, wie in Abbildung 11 dargestellt. Auswerter, die immer wieder eingesetzt werden, müssen nicht an jedem Training teilnehmen. Eine Begutachtung ihrer Leistung (siehe 5.1.3 Qualitätssicherung und Qualitätskontrolle) macht es einfacher zu entscheiden, ob ein Auswerter zu einer Schulung gebeten wird, eine Auffrischung benötigt oder ersetzt werden soll.

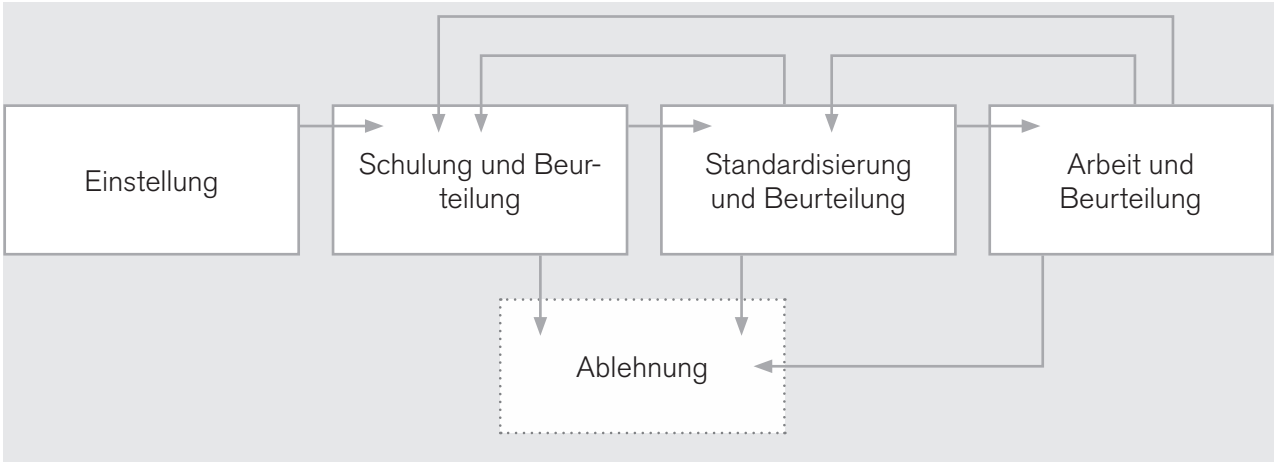


Abb. 11: Einstellung, Schulung und Beurteilung von Auswertern/Bewertern

Hilfestellung für die Beurteilung von Teilnehmerleistungen

Ein standardisierter Lösungsschlüssel ist die beste Möglichkeit, um den Auswertern die richtigen Lösungen zur Verfügung zu stellen. Lösungsschlüssel werden zur gleichen Zeit erstellt wie die Items selbst und unterliegen dem gleichen Redaktionsprozess. Der Schlüssel sollte die möglichen Antworten so umfangreich wie möglich abdecken und muss für die Auswerter eindeutig sein.

Abbildung 12 zeigt ein Beispiel, in dem die Teilnehmenden aufgefordert werden, die Lücke zu füllen. Der Schlüssel gibt mögliche Antworten. Die Gesamtzahl der zu erreichenden Punkte für dieses Item beträgt zwei. Je nach Grad der Aufgabenerfüllung werden beide möglichen Punkt oder nur ein Punkt vergeben.

Ein klares Layout des Lösungsschlüssels und anderer Unterlagen gewährleistet eine effiziente, korrekte und reliable Arbeit der Auswerter.

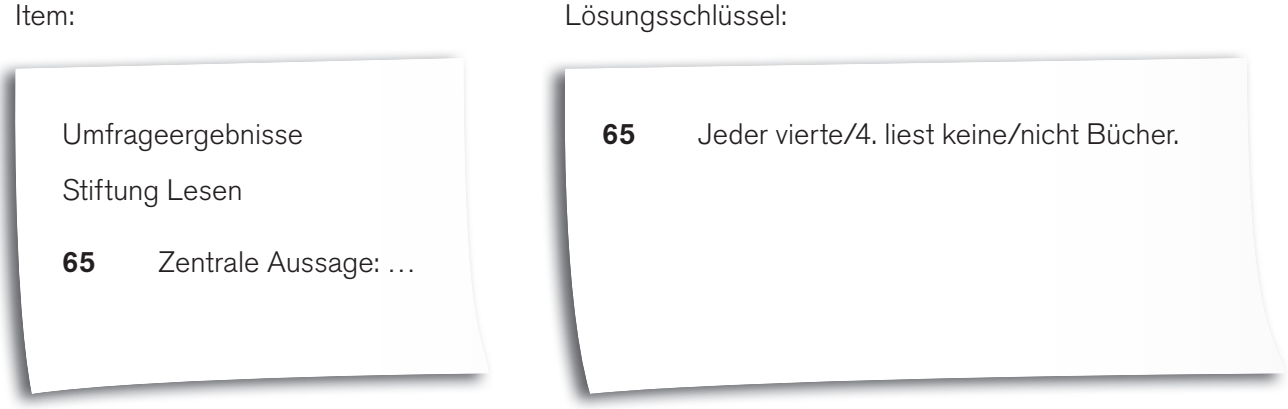


Abb. 12: Beispiel für ein Lückentext-Item

Weitere richtige Lösungen, die nicht im Lösungsschlüssel enthalten sind, wären ebenfalls möglich. Daher sollten Auswerter aufgefordert werden, alle Alternativantworten, die sie für richtig halten, zu notieren. Wenn sie tatsächlich als richtig beurteilt werden, müssen die Teilnehmenden auch hierfür einen Punkt erhalten. Bei einem kleinen Kreis von Auswertern reichen regelmäßige Gespräche mit dem Testentwickler aus, um Fragen dieser Art unmittelbar zu klären. Wenn aber im anderen Fall der Lösungsschlüssel überarbeitet und geändert wird, müssen einige oder alle Antwortbögen noch einmal ausgewertet werden.

Organisation des Auswertungsprozesses

Die Auswertungszeit ist im Allgemeinen begrenzt, da die Ergebnisse den Prüfungsteilnehmerinnen und -teilnehmern innerhalb einer bestimmten Frist mitgeteilt werden müssen. Der notwendige Zeiteinsatz kann anhand der Anzahl der Teilnehmenden und der Auswerter geschätzt werden. Wenn möglich, sollte die benötigte Zeit bzw. die Zahl der Auswerterinnen und Auswerter eher großzügig angesetzt werden, so dass auch eventuelle Schwierigkeiten bewältigt werden können.

Bei einer großen Anzahl von Teilnehmenden und Auswertern wird ein System benötigt, das den Verbleib der Antwortbögen im gesamten Prozess nachvollziehbar macht. Ein einfaches System besteht darin, die Antwortbögen zu nummerieren, mit der Nummer des Auswerters, dem Eingangsdatum und dem Auswertungsdatum zu versehen und diese Daten zu erfassen. Dies ermöglicht es dem Testanbieter, die Auswertungszeit und die benötigte Anzahl von Auswertern für eine bestimmte Anzahl von Teilnehmenden einzuschätzen.

Ein solches System kann auch dazu benutzt werden, die Leistung eines einzelnen Auswerters zu kontrollieren, z.B. die zur Auswertung eines Testteils durchschnittlich benötigte Zeit. Auch die durchschnittliche Anzahl der fehlerhaften Auswertungen kann bei einer Überprüfung erfasst werden. Für statistische Zwecke ist es meistens ausreichend, die Arbeit der Auswerter stichprobenartig zu kontrollieren.

5.1.2 Maschinelle Auswertung

Maschinelle Auswertung setzt normalerweise einen **optischen Belegleser** bzw. eine **OMR-**Technologie (OMR = *Optical Mark Reading/Optical Mark Recognition*) voraus. Ein solches OMR-System ist vor allem sinnvoll, wenn sehr viele Tests ausgewertet werden müssen und die Aufgabentypen keine Beurteilung durch Menschen erfordern (z.B. *Multiple Choice*-, Richtig/Falsch- oder Zuordnungsaufgaben). Teilnehmende können ihre Lösungen auf einen speziell entwickelten Bogen zur optischen Markierungserkennung notieren, wie in Abb. 13 abgebildet. Dieser wird in einen OMR-Scanner eingelesen, so dass das Ergebnis computergesteuert erfasst und gespeichert werden kann. OMR-Technologie wird auch für Items genutzt, die manuelle Auswertung oder Bewertung erfordern. Dann erfasst man die manuellen Auswertungen auf dem OMR-Bogen und scannt sie ein.

Scanner beschleunigen den Auswertungsprozess und reduzieren menschliche Fehler. Allerdings kann es vorkommen, dass der Scanner Teilnehmer-Markierungen nicht erfasst oder nicht beabsichtige Markierungen aufzeichnet. Daher müssen Plausibilitätsprüfungen durchgeführt werden, um solche Fehler zu vermeiden. So müssen alle OMR-Bögen dahingehend überprüft werden, ob sie Angaben erhalten, die für die Lösung nicht vorgesehen waren – beispielsweise mehrere Markierungen, obwohl das Item nur eine Lösung zulässt. Korrekturen auf dem OMR-Bogen müssen per Hand ausgeführt werden.

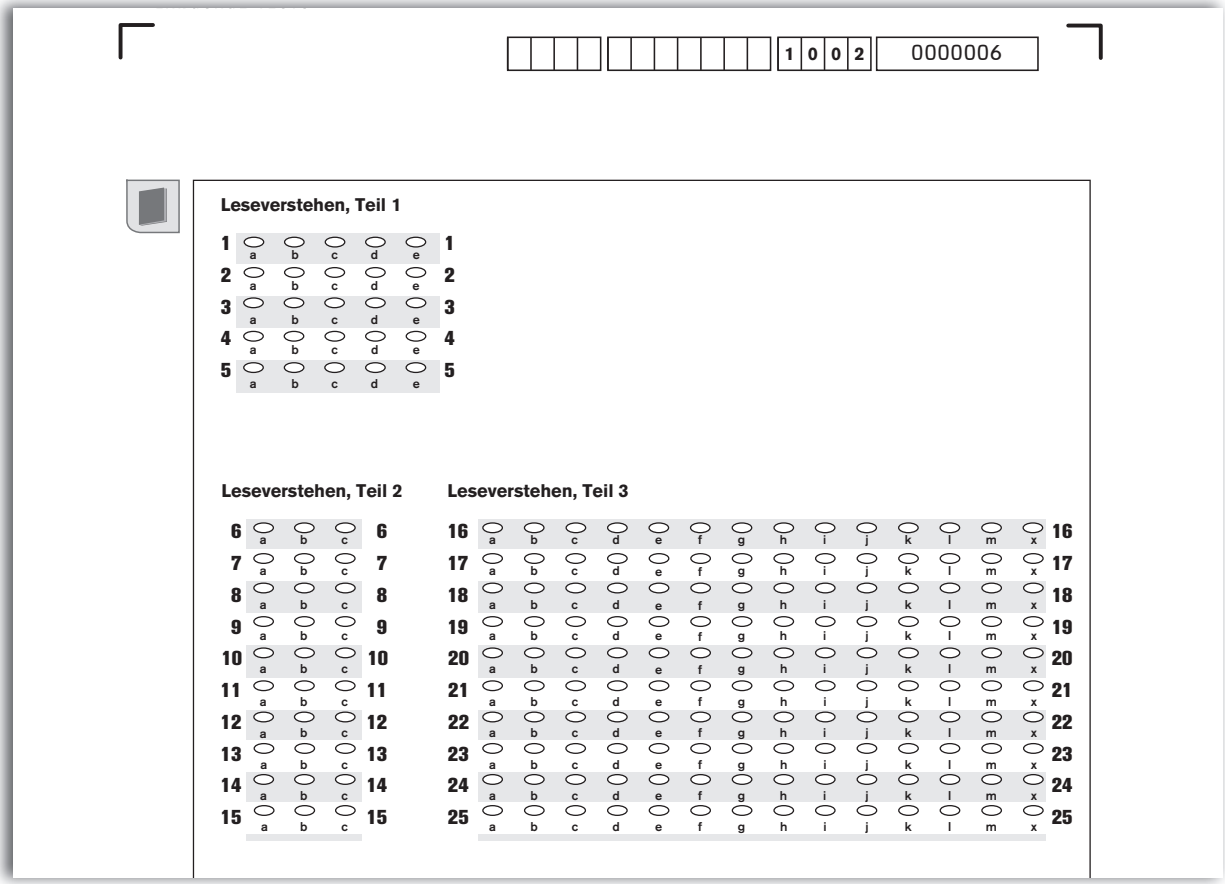


Abb. 13: Teil eines OMR-Bogens

5.1.3 Bewertung

Die Begriffe „**Bewertung**“ und „**Bewerter**“ werden hier für den Kontext benutzt, in dem eine qualifizierte Beurteilung in viel stärkerem Maße als beim einfachen Auswerten notwendig ist. Für eine solche Beurteilung kann der Testanbieter nicht eindeutig eine einzige korrekte Antwort im Vorfeld vorgeben. Daher gibt es hier mehr Raum für unterschiedliche Einschätzungen als bei einer anderen Form der Ergebnisfindung und somit eine größere Gefahr für Uneinheitlichkeit zwischen den Bewertern oder im individuellen Bewertungsprozess. Schulung, Kontrolle und korrigierende Rückmeldungen stellen insgesamt die Verlässlichkeit und Genauigkeit der Bewertung sicher.

Vieles von dem, was über manuelle Auswertung gesagt wurde, gilt auch für die Bewertung: Die Organisation muss eine effiziente Nutzung der Ressourcen ermöglichen und durch ständige Qualitätskontrollen die Genauigkeit der Bewertung sicherstellen. Die Reliabilität der Bewertungen muss ebenfalls überprüft werden (siehe Kapitel 1.3, Anhang VII).

Bewertungsskalen

Bewertungskompetenz beruht in den meisten Fällen auf einer **Bewertungsskala**. Diese besteht aus Deskriptoren, die die Leistung auf den verschiedenen Niveaus beschreiben und zeigen, welche Bewertung jedes Leistungsniveau erhalten soll.

Bewertungsskalen reduzieren die Variationsbreite, die der Beurteilung durch Menschen eigen ist. Es gibt verschiedene Möglichkeiten:

- Holistische (ganzheitliche) oder analytische Skalen:
  Eine Teilnehmerleistung kann aufgrund einer einzelnen **Skala** durch eine einzige Note oder eine andere Form der Einordnung bewertet werden. In diesem Fall beschreibt die Bewertungsskala jedes Leistungsniveau, ggf. als eine Reihe von Eigenschaften. Die Bewerterin oder der Bewerter sucht das Leistungsniveau, das die Leistung am besten beschreibt. Alternativ können Skalen entwickelt werden, die mehrere Kriterien definieren (z. B. die kommunikative Wirkung, Genauigkeit, die Abdeckung des erwarteten Inhalts etc.), und für jedes dieser Kriterien wird eine eigene Bewertung vergeben. Beide Ansätze mögen sich auf eine ähnliche Idee von Sprachkompetenz beziehen, die in ähnlichen Worten beschrieben wird – der Unterschied liegt in der erforderlichen Art der Beurteilung.

- Relative oder absolute Skalen:
  Skalen können entweder mithilfe relativer Bewertungen formuliert sein (mit Bezeichnungen wie „mangelhaft“, „befriedigend“, „gut“) oder können darauf abzielen, die Leistungsniveaus mit positiven, klaren Begriffen zu definieren. Wenn wir die Leistung gemäß den GER-Kompetenzskalen und -stufen interpretieren wollen, so erscheint die zweite Möglichkeit besser, und die GER-Deskriptorenskalen sind eine gute Quelle für die Entwicklung solcher Bewertungsskalen.

- Skalen oder Checklisten:
  Eine Alternative zu den oben beschriebenen Bewertungsskalen ist die Punktevergabe nach einer Liste von Ja/Nein-Bewertungen, also die Beurteilung, ob eine Leistung bestimmte Anforderungen erfüllt oder nicht.

- Allgemeine oder aufgabenspezifische Skalen:
  Eine Prüfung kann eine allgemeine Skala für alle Aufgaben verwenden oder Bewertungskriterien aufstellen, die sich auf bestimmte Aufgaben beziehen. Eine Kombination ist ebenfalls möglich, z. B. können spezielle Kriterien für die Erfüllung einer Aufgabe gelten (eine Auflistung der abzudeckenden inhaltlichen Punkte), wogegen andere Skalen allgemein bleiben.

- Vergleichende oder absolute Beurteilung:
  Skalen können mithilfe von beispielhaften Teilnehmerleistungen aufgebaut werden, so dass der Bewerter das Leistungsniveau nicht absolut einschätzen muss, sondern lediglich entscheidet, ob es höher, niedriger oder gleichwertig in Bezug auf ein oder mehrere Beispiele ist. Die Bewertung besteht also in einer Rangordnung innerhalb der Skala. Die Interpretation dieser Einordnung, z. B. im Sinne der GER-Kompetenzstufen, hängt von der Beurteilung der durch die Teilnehmerbeispiele illustrierten Leistungsniveaus ab. Ein solcher Ansatz funktioniert sicherlich am besten bei aufgabenspezifischen Beispielen.

Obwohl zwischen diesen Ansätzen anscheinend große Unterschiede bestehen, unterliegen sie doch alle den gleichen Prinzipien:

- Alle Bewertungen hängen davon ab, dass die Bewerterin/der Bewerter die Kompetenzstufen versteht.
- Beispiele sind absolut notwendig, um dieses Verständnis zu fördern und zu kommunizieren.
- Bei der Arbeit mit Skalen sind die Aufgaben, die einer zu bewertenden Leistung zugrunde liegen, von großer Bedeutung.

Traditionell hatten Kompetenzstufen eine Bedeutung, die sich eher auf den Kontext einer bestimmten Prüfung und ihrer Teilnehmer bezog. Dadurch waren sie schwer mit den Kompetenzstufen für andere Prüfungen zu vergleichen. Die Entwicklung eines Kompetenzrahmens wie jene des GER bietet die Möglichkeit, die in einem bestimmten Kontext verwendeten Stufen auch auf andere Kontexte zu beziehen und interpretieren zu können. Das hat Folgen für die Art und Weise, wie die Bewertungsskalen formuliert sind.

Früher wurden die Kompetenzstufen implizit verstanden und die Skalen mit relativen, bewertenden Angaben ausgearbeitet. Heute dagegen werden die Skalen häufig gemäß dem GER-Ansatz aufgestellt und beschreiben die Leistungsniveaus mit klaren, positiven Aussagen, die einen hohen Wiedererkennungswert haben. Dennoch bleiben Beispiele (mehr noch als die Deskriptoren) ein absolutes Muss, um das Niveau zu definieren und zu kommunizieren. Testanbieter werden auf diese Weise dazu ermutigt, deutlicher auszudrücken, was es bedeutet, eine bestimmte Kompetenzstufe zu erreichen.

Der GER regt dazu an, in Richtung auf eine kriterienorientierte Bestimmung von Kompetenzstufen zu denken und zu arbeiten. Zwei Aspekte müssen beim Definieren einer Kompetenzstufe berücksichtigt werden: *Was* kann jemand und *wie* gut kann er es? Das „Was“ wird in der Prüfung durch die spezifischen Aufgaben festgelegt. Wie gut eine Aufgabe bewältigt wurde, haben die Bewerterinnen und Bewerter zu beurteilen.

Daher funktioniert auch der traditionell bewertende Ansatz ausreichend gut, so lange die Aufgaben richtig gewählt wurden und sich die Beurteilung auf die Leistung in diesen Aufgaben bezieht. Die Aufgaben sind also von allergrößter Bedeutung, wenn mit Skalen gearbeitet wird, auch wenn bei der Definition einer „bestandenen“ Leistung mehr oder weniger direkter Bezug auf die Aufgaben genommen wird.

Der GER beschreibt Aspekte subjektiver Beurteilung (Kapitel 9.3.8).

Der Bewertungsprozess

Damit die Bewertung gut vonstatten geht, müssen alle Bewerterinnen und Bewerter den zugrunde liegenden Standard verstehen. Dieses gemeinsame Verständnis wird durch gemeinsam verwendete Beispiele von Teilnehmerleistungen entwickelt.

Bei Prüfungen von kleinerer Reichweite können sich die Bewerter auf ein gemeinsames Verständnis einigen, indem sie offen und gleichberechtigt diskutieren. Dies bedeutet zwar, dass alle Teilnehmenden gleich

behandelt werden, was aber nicht garantiert, dass die vereinbarten Standards über die jeweilige Testsituation vor Ort hinaus verstanden werden oder auf mehrere Prüfungen angewendet werden können. Bei Prüfungen mit größerer Reichweite müssen die Standards beständig und aussagekräftig sein. Hierzu sind erfahrene Bewerterinnen und Bewerter notwendig, die neuen Bewertern ihre Standards kommunizieren können.

Ein Kreis von erfahrenen Bewertern ist also gewöhnlich das zentrale Element zur Aufrechterhaltung von Spandards durch Qualifizierung, Qualitätssicherung und Korrektur der anderen Bewerter.

Solch ein hierarchisches System umfasst verschiedene Stufen, wie in Abbildung 14 gezeigt. Dies kann eine effiziente Art sein, Qualifizierungsseminare zu organisieren oder die Arbeit der Bewerter zu kontrollieren. Bis zu einem gewissen Grad reduzieren aber moderne Informationstechnologien und die Entwicklung von webbasierten Trainingsangeboten die Notwendigkeit einer solchen Hierarchie. Für jede Stufe der Bewertung und Bewerterqualifizierung ist es wichtig, dass dieselben beispielhaften Teilnehmerleistungen mit Musterbewertung die korrekte Weitergabe der Standards sicherstellen.

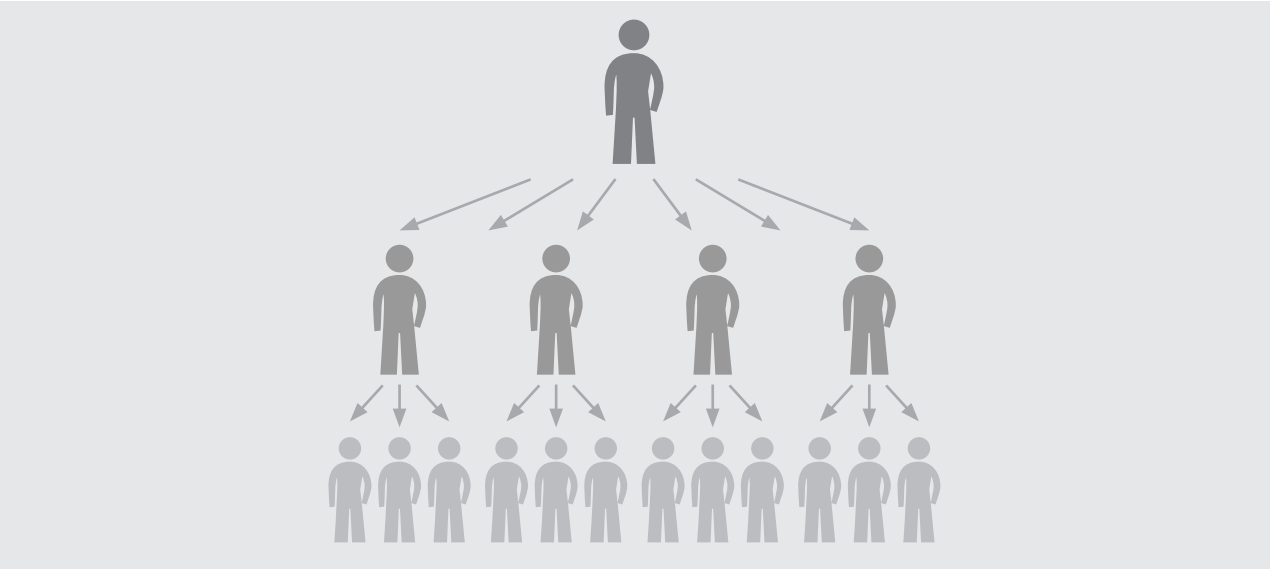


Abb. 14: Aufrechterhaltung der Standards durch ein Teamleiter-System

Bewerterqualifizierung

Ziel der Schulungen von Bewerterinnen und Bewertern sind einheitliche und genaue Bewertungen. Standardisierung bedeutet, dass alle Bewerter die angestrebten Standards anwenden. Mit dem GER als Bezugspunkt für solche Standards sollte Qualifizierung also damit beginnen, die Bewerter mit dem GER vertraut zu machen und ihnen Beispiele zur Veranschaulichung der Leistungen beim Sprechen und Schreiben zu geben (Europarat 2009). Es kann zudem notwendig sein, Bewertern ihre bereits vertrauten Bewertungsraster „abzutrainieren“. Der Qualifizierungsprozess umfasst insgesamt mehrere Schritte von der offenen Diskussion hin zu unabhängiger Bewertung von Teilnehmerleistungen aus dem betreffenden Test:

- gelenkte Diskussion eines Teilnehmerbeispiels, durch das die Bewerter die jeweilige Kompetenzstufe verstehen lernen
- unabhängige Bewertung eines Beispiels, gefolgt von einem Vergleich mit der Musterbewertung und Diskussion über die Gründe für abweichende Bewertungen
- unabhängige Bewertung mehrerer Beispiele, um zu erkennen, wie nah die Bewerter nun den Musterbewertungen kommen

Qualitätssicherung

Im besten Fall endet die Qualifizierungsphase damit, dass alle Bewerterinnen und Bewerter ausreichend einheitlich und genau arbeiten, so dass weiteres Feedback oder Korrekturen nicht notwendig sind. So läuft der Bewertungsprozess mit größtmöglicher Effizienz ab. Dennoch ist Qualitätssicherung notwendig, damit Probleme rechtzeitig erkannt und behoben werden können.

Es gibt vier Arten von Problemen oder so genannten „Bewerter-Effekten“:

1. Strenge oder Milde:
   Die Bewerterin oder der Bewerter bewertet systematisch zu hoch oder zu niedrig.

2. Umgang mit der Bewertungsspanne (Zentraltendenz):
   Der Bewerter nutzt ein zu kleines Spektrum und unterscheidet nicht klar genug zwischen starken und schwachen Leistungen.

3. Halo-Effekt:
   Wenn ein Bewerter mehrere Bewertungen abgeben muss, beeinflusst der erste Eindruck von dem Prüfungsteilnehmer auch die weiteren Bewertungen, unabhängig davon, wie die tatsächliche Leistung jeweils ausfällt.

4. Uneinheitlichkeit:
   Der Bewerter wendet die Standards uneinheitlich an, so dass seine Bewertungen nicht mit denen der anderen Bewertenden übereinstimmen.

Wie schwerwiegend diese Probleme sind, hängt zum Teil davon ab, welche Korrekturen möglich sind. Bezüglich der Strenge z. B. scheinen viele Bewerterinnen und Bewerter ein internalisiertes Maß zu besitzen. Versuche, dies zu vereinheitlichen, können das Gegenteil bewirken: Der Bewerter verliert das Vertrauen in sich selbst und bewertet nicht mehr konsequent. Daher ist es oft besser, ein gewisses Maß an systematischer Strenge oder Milde zu akzeptieren, so lange dies mit einem statistischen Verfahren ausgeglichen werden kann. *Scaling* oder ein *Item Response*-Modell sind zwei geeignete Optionen (vgl. Appendix VIII).

Die Verwendung einer zu kleinen Bewertungsspanne kann nur zum Teil statistisch ausgeglichen werden. Unheitlichkeit ist überhaupt nicht statistisch korrigierbar. Daher müssen diese beiden Probleme erkannt und behoben werden, indem die Bewerter entweder noch einmal geschult oder ersetzt werden.

Ein gewisses Qualitätssicherungssystem ist also notwendig. Eine Überprüfung ist relativ einfach, wenn bei der Bewertung eine Schreibleistung von einem Bewerter zum anderen weitergegeben wird. Mündliche Bewertung ist sehr viel schwieriger zu kontrollieren, es sei denn, die Leistung wird aufgezeichnet. Hier sollten also Bewerterinnen und Bewerter vor der Prüfung besonders gründlich geschult und richtig eingeschätzt werden. Statistiken über die Leistungen des Prüfers können bei diesem Prozess hilfreich sein (siehe Anhang VII).

Der Qualitätssicherungsprozess reicht von kleinen bis hin zu umfangreichen Maßnahmen. Einfache Maßnahmen sind beispielsweise informelle Stichproben und mündliches Feedback an den Bewerter. Umfangreiche Maßnahmen können die Korrektur einer größeren Stichprobe von Bewertungen und statistische Aufzeichnungen über die Bewerterleistung sein. Eine beliebte Methode besteht darin, dem Bewerter einige bereits bewertete Teilnehmerleistungen unter die noch zu bewertenden Leistungen zu geben und zu überprüfen, wie eng die Bewertungen beieinander liegen. Allerdings dürfen die bereits bewerteten Teilnehmerleistungen für eine verlässliche Analyse nicht von anderen zu unterscheiden sein, sodass beispielsweise keine Fotokopien benutzt werden können. Diese Methode eignet sich also nur bei computergestützten Prüfungen oder wenn Teilnehmerleistungen auf Papier in ein Online-Korrektursystem eingescannt werden.

Eine andere Möglichkeit, Fehlentscheidungen der Bewerterinnen und Bewerter zu minimieren und einzelne Bewertungen miteinander zu vergleichen (was zur Erkennung und statistischen Korrektur einiger Bewerter-Effekte führt), ist die durchgehende **Doppelbewertung** oder die Mehrfachbewertung einer Teilmenge, bei der nur ein gewisser Prozentsatz der Teilnehmerleistungen mehrfach bewertet wird. Je nach statistischer Methode müssen die Informationen auf die eine oder andere Art kombiniert werden und im Endeffekt für die Teilnehmenden zu einer Endbewertung führen.

5.2 Benotung

Der gesamte Prozess der Konzeption, Entwicklung, Durchführung und Bewertung eines Tests, so wie er bis hierher beschrieben wurde, führt zu dem Punkt, an dem schließlich die Leistung der Teilnehmenden beurteilt und bekannt gegeben werden kann.

In einigen Fällen werden Prüfungsteilnehmerinnen und teilnehmer einfach nach ihrer Leistung sortiert, von sehr gut bis sehr schlecht. Dann werden vielleicht beliebige Grenzen gezogen, um sie in Gruppen zu unterteilen – z. B. die besten 10 % erhalten die **Note** 1, die nächsten 30 % erhalten die Note 2 usw. Dieser *normbezogene* Ansatz kann zwar innerhalb der Gesellschaft wichtige Zwecke erfüllen, ansonsten ist er aber ungeeignet, da die Leistung nur im Bezug auf die anderen Teilnehmenden bewertet wird. Er sagt nichts darüber aus, was Leistung tatsächlich bedeutet, etwa mit Bezug auf Sprachkompetenzniveaus.

Ein anderer, aussagekräftigerer Ansatz ist der *kriteriumbasierte*, bei dem die Leistung in Bezug auf festgelegte, absolute Kriterien oder Standards beurteilt wird. Dies ist ganz klar der Fall bei Prüfungen, die die Ergebnisse in Bezug auf GER-Stufen angeben.

Eine Prüfung kann so ausgelegt sein, dass sie sich über mehrere GER-Stufen erstreckt oder sich nur auf eine bezieht. Im zweiten Fall haben die Prüfungsteilnehmerinnen und teilnehmer, die die Stufe erreicht haben, „bestanden" und die anderen sind „durchgefallen". Bestehen oder Durchfallen kann je nach Leistung differenziert ausgewiesen werden.

Die Festlegung einer Punktzahl, mit der ein bestimmtes Niveau erreicht wird, wird **Standardsetzung** (oder *Standard Setting*) genannt. Hierfür ist eine subjektive Einschätzung unvermeidlich, die so weit wie möglich auf empirischen Daten basieren muss.

Produktive Fertigkeiten (Sprechen, Schreiben) verlangen andere Methoden der Standardsetzung als rezeptive Fertigkeiten (Lesen, Hören), die meist **objektiv ausgewertet** werden. Den produktiven Leistungen kann man sich leichter annähern. Beim Lesen und Hören ist es schwieriger, da wir mentale Vorgänge interpretieren müssen, die nur indirekt beobachtbar sind. Daher ist die Kompetenzstufe schwerer in Kriterien zu fassen.

Für einen Test, der verschiedene Fertigkeiten abprüft, müssen die Standards für jede Fertigkeit einzeln festgelegt werden. Das führt zu der Frage, wie diese Testergebnisse zusammengeführt werden können (siehe Kapitel 5.3 für mehr Informationen).

An dieser Stelle sei auf das *Manual for Relating Language Examinations to the CEFR* (Council of Europe 2009) hingewiesen, welches detaillierte Informationen zur Standardsetzung gibt. Zu beachten ist im Hinblick auf Aufbau und Terminologie des o.g. Handbuchs:

- Kapitel 6 – „Standard Setting Procedures" – bezieht sich ausschließlich auf objektiv zu bewertende Testaufgaben, also etwa im Lesen und Hören.

- Produktive Fertigkeiten werden unter der Überschrift „Standardisation Training and Benchmarking“ in Kapitel 5 beschrieben.
- Kapitel 7 zur **Validierung** sollte ebenfalls aufmerksam gelesen werden. Hier geht es um zwei Ansätze zur Standardsetzung: aufgabenbasiert und lernerbasiert. Ein aufgabenbasierter Ansatz beruht auf Expertenurteilen über die Testitems, was in Kapitel 6 behandelt wird. Ein lernerbasierter Ansatz versucht über den Test hinaus zusätzliche Informationen über die Lernenden zu sammeln. Dies wird in Kapitel 7 erläutert.
- Dieser Aufbau heißt nicht, dass die aufgabenbasierte Standardsetzung wichtiger ist als der lernerbasierte Ansatz.

Streng genommen sollte die Standardsetzung nur einmal geschehen, nämlich zu dem Zeitpunkt, an dem die Prüfung zum ersten Mal durchgeführt wird. Jedoch zeigt die Praxis, dass es sich hierbei um einen iterativen Prozess handelt. Langfristig sollte Benotung allerdings die *Aufrechterhaltung* der Standards und nicht deren Festlegung bedeuten. Hierzu müssen angemessene Verfahrensweisen während des gesamten Prüfungsverfahrens existieren. Diese werden in den ergänzenden Dokumenten zum sog. *Manual* aufgeführt (North und Jones 2009).

5.3 Übermittlung der Ergebnisse

Der Testanbieter muss entscheiden, ob den Teilnehmenden ein einziges Ergebnis oder ein Ergebnisprofil übermittelt wird, das die Leistungen in jedem Testteil bzw. Subtest zeigt.

Ersteres ist häufiger anzutreffen, da die meisten Beteiligten die einfache Darstellungsform einer komplexen vorzuziehen scheinen. Die zweite Möglichkeit bietet einen informativeren Ansatz, der in einigen Situationen zweckdienlicher sein kann.

Eine dritte Möglichkeit wäre, beides anzubieten. Der GER betont, wie wichtig es ist, nach Möglichkeit Ergebnisprofile zu verwenden.

Wenn ein einziges Ergebnis benötigt wird, muss eine Methode gefunden werden, um die Bewertungen der einzelnen Teilfertigkeiten zusammenzuführen. Der Testanbieter muss entscheiden, wie jeder einzelne Teil **gewichtet** wird: alle Fertigkeiten gleich oder einige stärker als andere. Dies erfordert ggf. Änderungen bei den **Rohwerten** für die einzelnen Testteile (siehe Anhang VII).

Wenn ein Zertifikat überreicht wird, muss der Testanbieter Folgendes beachten:

- ob zusätzliches Material (z. B. die Kann-Beschreibungen) mitgeliefert wird, um die Bedeutung des Niveaus zu veranschaulichen,
- ob garantiert werden soll, dass es sich um ein Originalzertifikat handelt (indem Fälschung oder Veränderung erschwert oder eine Überprüfungsleistung angeboten wird),
- ob – und wenn ja, welche – Warnungen gegen bestimmte Interpretationen der Ergebnisse ausgesprochen werden sollten.

5.4 Schlüsselfragen

- Wie viel manuelle Auswertung ist nötig und mit welcher Häufigkeit?
- Wie aufwendig und wie oft wird bewertet?
- Welches Fachwissen brauchen Bewerterinnen und Bewerter?

- Wie wird die genaue und verlässliche Auswertung und Bewertung sichergestellt?
- Was ist der beste Weg, die Prüfungsteilnehmer im jeweiligen Prüfungskontext zu klassifizieren?
- Wem werden die Ergebnisse übermittelt und wie wird dies getan?

5.5 Weiterführende Literatur

Siehe ALTE (2 206) für eine Checkliste zur Selbsteinschätzung für Auswertung, Bewertung und Ergebniserstellung.

Kaftandjieva (2004), North und Jones (2009) und Figueras und Noijons (2009) geben Informationen zur Standardsetzung.

6 Qualitätssicherung

Es ist wichtig, die geleistete Arbeit zu Entwicklung und Einsatz des Tests zu überprüfen. Entspricht der Test einem akzeptablen Standard, oder müssen Änderungen vorgenommen werden? Ziel der Qualitätssicherung ist es festzustellen, ob während und unmittelbar nach der Durchführung der Prüfung alles korrekt abgelaufen ist. Erforderliche Änderungen können oft schnell durchgeführt werden. Verbesserungen kommen den aktuellen und den zukünftigen Prüfungsteilnehmerinnen und -teilnehmern zugute.

Die Evaluation des Tests ist ein komplexerer Vorgang, bei dem viele verschiedene Aspekte berücksichtigt werden. Hier geht man bis zur Testentwicklung zurück, bis hin zu den grundlegenden Fragen wie „Wird diese Prüfung wirklich benötigt?“, „Für welchen Zweck?“, „Für wen?“ und „Was versuchen wir zu prüfen?“. Dies ähnelt der Testentwicklungsphase, aber mit dem Vorteil, dass Daten und Erfahrungswerte von vorangegangenen Prüfungsereignissen vorliegen. Aufgrund des Umfangs und der Bedeutung kann diese Evaluation nicht Teil der normalen Testdurchführung sein und nicht nach jeder Prüfung erfolgen.

6.1 Routinemäßige Qualitätssicherung

Qualitätssicherung ist routinemäßig Teil der Testerstellung und Prüfungsdurchführung. Die Informationen hieraus werden genutzt, um sicherzustellen, dass alles, was mit der aktuellen Prüfungsdurchführung zusammenhängt, korrekt verläuft: Materialien werden regelgerecht erstellt, so dass sie pünktlich ausgeliefert werden können, Teilnehmende erhalten die korrekten Bewertungen etc. Weiterhin kann man dieselben Informationen nutzen, um allgemein die Effizienz der Prozesse zur Itemerstellung, Redaktionsarbeit, Versionserstellung, Bewertung etc. einzuschätzen. Diese Informationen kann wiederum für die Validitätsargumentation wichtig sein (siehe Anhang I), was man bei ihrer Sichtung gleich berücksichtigen sollte.

Dieses Handbuch hat bereits einige Beispiele für das Vorgehen im Qualitätssicherungsprozess aufgezeigt. Dazu gehören:

- Einholen von Expertenurteilen und Erprobungen, um sicherzustellen, dass Items gut erstellt sind (siehe Kapitel 3.4)
- Analyse der Teilnehmer-Lösungen, um zu entscheiden, ob die Items gut funktionieren (siehe Anhang VII)
- Einholen von Kommentaren, um zu sehen, wie gut die Organisation war (siehe Anhang VI)
- Sammlung und Analyse von Daten zur Testauswertung (siehe Anhang VII)
- Auch ist es durchaus sinnvoll, die Effizienz der Arbeit zu überwachen. Testanbieter können messen, wie lange die jeweiligen Phasen dauern, und entscheiden, ob zu viel oder zu wenig Zeit angesetzt wurde.

6.2 Periodische Evaluation der Prüfung

Eine grundlegendere Evaluation findet gelegentlich jenseits der regulären Qualitätssicherung statt. Dies kann in regelmäßigen Abständen geschehen oder jedenfalls bei wichtigen Änderungen, also z. B. einer anderen Zielgruppe, einer neuen Verwendung des Tests, einem neuen Lehrplan. Auch die routinemäßige Qualitätssicherung kann die Notwendigkeit einer umfassenderen Revision aufzeigen. Eine Evaluation ermöglicht in jedem Fall die detaillierte Begutachtung des Tests und der Art, wie er erstellt wird. Informatio-

nen aus der Prüfungspraxis, z. B. aus der Überwachung des Bewerter-Verhaltens, können für die Evaluation von Nutzen sein. Zusätzlich können Testanbieter entscheiden, dass weitere Informationen benötigt werden, die speziell für die Evaluation eingeholt werden müssen.

Für die Evaluation werden Informationen eingeholt und festgehalten, die bei der Entscheidung darüber helfen, welche Aspekte des Tests überarbeitet werden müssen (z. B. die Zusammensetzung, das Format, die Durchführungsregeln). Durchaus möglich ist das Ergebnis, dass nur sehr wenige oder gar keine Änderungen vorgenommen werden müssen.

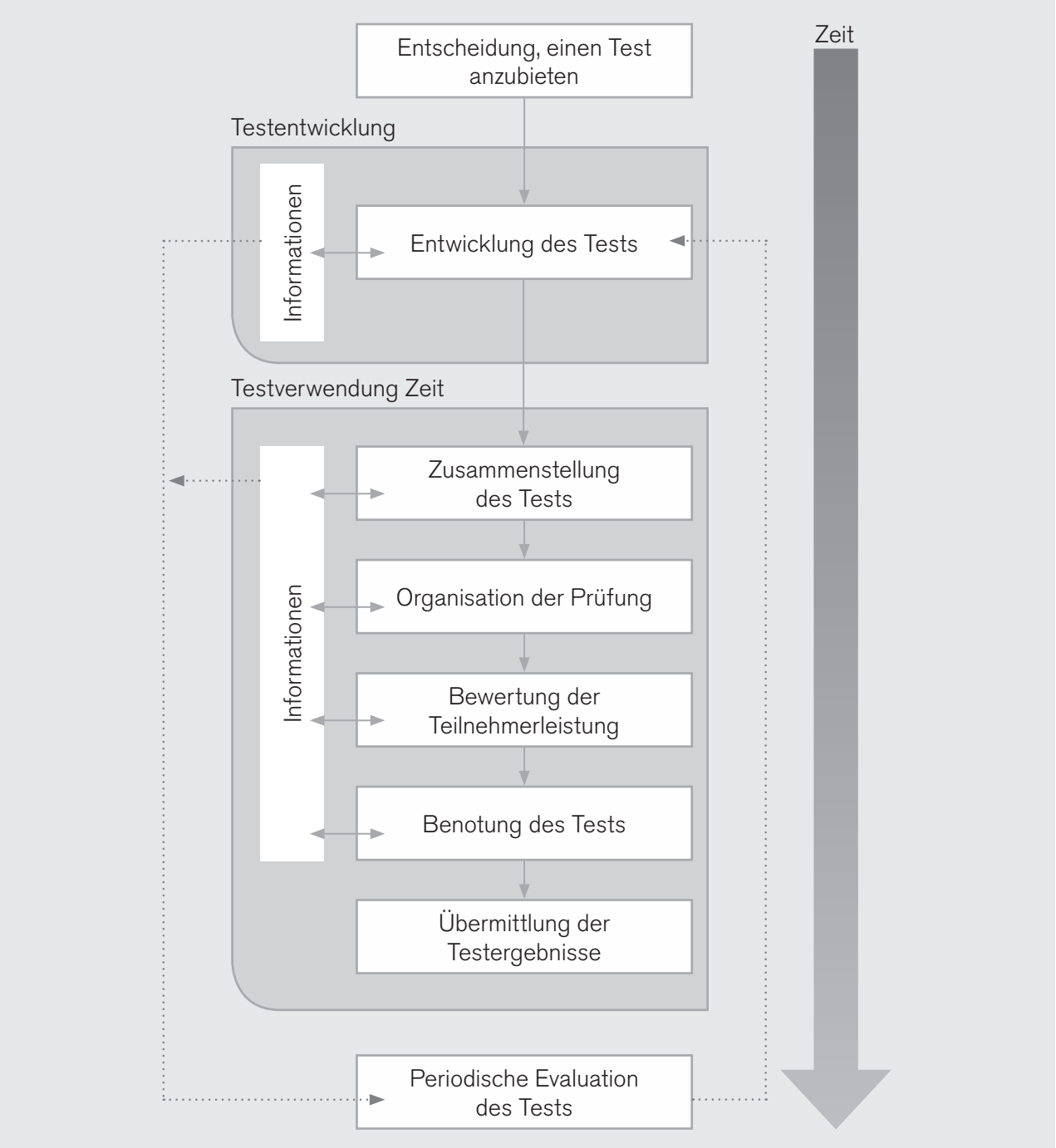


Abb. 15: Der allgemeine Testzyklus und die periodische Evaluation

Abbildung 15 ist eine Kopie von Abbildung 5 (Kapitel 1.5.1) mit dem Zusatz der regelmäßigen Evaluation. Sie zeigt, dass die Ergebnisse dieser Nachbearbeitung auf die erste Phase des Prozesses zurückwirken: die Entscheidung, einen Test anzubieten. Der Prozess der Testentwicklung wird durch die Evaluation noch einmal durchlaufen.

6.3 Bereiche der Qualitätssicherung

Arbeiten zur Qualitätssicherung gehören zu den Routinearbeiten bei der Testentwicklung und -durchführung. Sie zeigen dem Testanbieter, ob alles so funktioniert, wie es sollte, oder was andernfalls zu ändern ist. Qualitätssichernde Maßnahmen können auch anderen, wie etwa Schulen oder Akkreditierungsgremien, zeigen, dass sie der Prüfung vertrauen können. Aus beiden Perspektiven kommt die Überprüfung dessen, was gemacht wird und ob es gut genug gemacht wird, weitgehend einer Auditierung der Prüfungsvalidität gleich.

ALTE (2007) hat eine Auflistung mit 17 Kernpunkten aufgestellt, den Mindeststandards, die es Testanbietern ermöglichen, einen Validitätsbeleg aufzubauen. Sie sind in die folgenden fünf Bereiche gegliedert:

- Prüfungsentwicklung
- Durchführung und Logistik
- Bewertung und Benotung
- Analyse der Ergebnisse
- Kommunikation mit Beteiligten

Diese Mindeststandards sollen zusammen mit ausführlicheren und genaueren Auflistungen verwendet werden, wie z. B. *ALTE Content Analysis Checklists* (ALTE 2004a–k, 2005, 2006a–c).

Auch andere Handreichungen können den Testanbietern den Aufbau und die Prüfung ihrer Validitätsargumentation erleichtern. Jones, Smith und Talley (2006: 490–2) geben eine Liste mit 31 Kernpunkten für Tests mit kleinerem Umfang. Viele ihrer Punkte basieren auf den *Standards for Educational and Psychological Testing* (AREA et al 1999).

6.4 Schlüsselfragen

- Welche Daten müssen zur effizienten Qualitätssicherung der Prüfung gesammelt werden?
- Werden einige dieser Daten bereits während der Prüfungsdurchführung gesammelt, um Routineentscheidungen zu treffen? Wie können diese auf einfache Art und Weise für beide Zwecke genutzt werden?
- Können die Daten aufbewahrt und später bei der Evaluation verwendet werden?
- Wer soll bei der Evaluation involviert sein?
- Welche Ressourcen stehen für die Evaluation zur Verfügung?
- Wie oft sollte eine Evaluation stattfinden?
- Können einige Punkte aus der o.g. Liste nützlich bei der Überprüfung der Validitätsargumentation sein?

6.5 Weiterführende Literatur

ALTE (2007) gibt verschiedene Kategorien für die Überprüfung eines Tests an.

Siehe ALTE (2002) für eine Checkliste zur Selbsteinschätzung für Testanalyse und Nachbereitung.

Fulcher und Davidson (2009) zeigen einen interessanten Weg auf, wie die erhobenen Daten zur Prüfungsrevision genutzt werden können. Sie bedienen sich der Metapher eines Gebäudes, um die Teile des Tests zu zeigen, die regelmäßig und weniger regelmäßig geändert werden müssen.

Beschreibungen der verschiedenen Aspekte der Testrevision finden sich bei Weir und Milanovic (2003).

Literaturverzeichnis

AERA, APA, NCME (1999): *Standards for Educational and Psychological Testing*. Washington DC: AERA Publishing.

Alderson, J. C.; Clapha, C. und Wall, D. (1995): *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

ALTE (1994): *Code of Practice.* http://www.alte.org/attachments/files/code_practice_eng.pdf, abgerufen am 05.08.2012.

ALTE (2002): *ALTE Quality Management and Code of Practice Checklist: test analysis and post examination review.* http://www.alte.org/cop/copcheck.php, abgerufen am 07.12.2009.

ALTE (2004a): *Development and descriptive checklist for tasks and examinations: general.* http://www.alte.org/downloads/index.php, abgerufen am 07.12.2009.

ALTE (2004b): *Individual component checklist: reading.* http://www.alte.org/attachments/files/reading_check.pdf, abgerufen am 05.08.2012.

ALTE (2004c): *Individual component checklist: structural competence.* http://www.alte.org/attachments/files/structural_comp.pdf, abgerufen am 05.08.2012.

ALTE (2004d): *Individual component checklist: listening.* http://www.alte.org/attachments/files/listening_check.pdf, abgerufen am 05.08.2012.

ALTE (2004e): *Individual component checklist: writing.* http://www.alte.org/attachments/files/writing_check.pdf, abgerufen am 05.08.2012.

ALTE (2004f): *Individual component checklist: speaking.* http://www.alte.org/attachments/files/speaking_check.pdf, abgerufen am 05.08.2012.

ALTE (2004g): *Individual component checklist – for use with one task: reading.* http://www.alte.org/attachments/files/reading_check_onetask.pdf, abgerufen am 05.08.2012.

ALTE (2004h): *Individual component checklist – for use with one task: structural competence.* http://www.alte.org/attachments/files/structural_comp_onetask.pdf, abgerufen am 05.08.2012.

ALTE (2004i): *Individual component checklist – for use with one task: listening.* http://www.alte.org/attachments/files/listening_check_onetask.pdf, abgerufen am 05.08.2012.

ALTE (2004j): *Individual component checklist – for use with one task: writing.* http://www.alte.org/attachments/files/writing_check_onetask.pdf, abgerufen am 05.08.2012.

ALTE (2004k): *Individual component checklist – for use with one task: speaking.* http://www.alte.org/attachments/files/speaking_check_onetask.pdf, abgerufen am 05.08.2012.

ALTE (2005): *ALTE materials for the guidance of test item writers* (1995, neu Juli 2005). http://www.alte.org/attachments/files/item_writer_guidelines.pdf, abgerufen am 05.08.2012.

ALTE (2006a): *ALTE Quality Management and Code of Practice Checklist: test construction.* http://www.alte.org/cop/copcheck.php, abgerufen am 07.12.2009.

ALTE (2006b): *ALTE Quality Management and Code of Practice Checklist: administration and logistics.* http://www.alte.org/cop/copcheck.php, abgerufen am 07.12.2009.

ALTE (2006c): *ALTE Quality Management and Code of Practice Checklist: marking, grading, results.* http://www.alte.org/cop/copcheck.php, abgerufen am 07.12.2009.

ALTE (2007): *Minimum standards for establishing quality profiles in ALTE Examinations.* http://www.alte.org/downloads/index.php, abgerufen am 07.12.2009.

ALTE (2008a): *The ALTE Can Do Project.* http://www.alte.org/attachments/files/alte_cando.pdf, abgerufen am 05.08.2012.

ALTE (2008b): *ALTE Quality Management and Code of Practice Checklists.* http://www.alte.org/cop/copcheck.php, abgerufen am 07.12.2009.

ALTE Members (1998): *Multilingual glossary of language testing terms* (Studies in Language Testing volume 6), Cambridge: Cambridge University Press.

ALTE Members (2005a): *The CEFR Grid for Speaking, developed by ALTE Members (input) v. 1.0.* http://www.coe.int/T/DG4/Portfolio/documents/ALTE%20CEFR%20Speaking%20Grid%20INput51.pdf, abgerufen am 05.08.2012.

ALTE Members (2005b): *The CEFR Grid for Speaking, developed by ALTE Members (input) v. 1.0.* http://www.coe.int/T/DG4/Portfolio/documents/ALTE%20CEFR%20Speaking%20Grid%20UTput5l.pdf, abgerufen am 03.04.2009.

ALTE Members (2007a): *The CEFR Grid for Writing Tasks v. 3.7 (analysis).* http://www.coe.int/T/DG4/Portfolio/documents/CEFRWritingGridv3_i_analysis.doc, abgerufen am 03.04.2009.

ALTE Members (2007b): *The CEFR Grid for Writing Tasks v. 3.7 (Präsentation).* http://www.coe.int/T/DG4/Portfolio/documents/CEFRWritingGridv3_i_presentation.doc, abgerufen am 03.04.2009.

ALTE Working Group on Code of Practice (2001): *The Principles of Good Practice for ALTE Examinations.* http://www.alte.org/downloads/index.php, abgerufen am 07.12.2009.

Assessment Systems (2009): *Iteman 4.* Software. Assessment Systems.

Bachman, L. F. (1990): *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F. (2004): *Statistical Analysis for Language Assessment.* Cambridge: Cambridge University Press.

Bachman, L. F. (2005): *Building and supporting a case for test use.* In: Language Assessment Quarterly 2 (1), 1–34.

Bachman, L. F.; Black, P.; Frederiksen, J.; Gelman, A.; Glas, C. A. W.; Hunt, E.; McNamara, T. und Wagner, R. K. (2003): *Commentaries Constructing an Assessment Use Argument and Supporting Claims About Test Taker-Assessment Task Interactions in Evidence-Centered Assessment Design.* In: Measurement: Interdisciplinary Research & Perspective 1 (1), 63–91.

Bachman, L. F. und Palmer, A. S. (1996): *Language testing in practice: designing and developing useful language tests.* Oxford: Oxford University Press.

Bachman, L. F. und Palmer, A. S. (2010): *Language assessment in practice: developing language assessments and justifying their use in the real world.* Oxford: Oxford University Press.

Banerjee, J. (2004): *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section D: Qualitative Analysis Methods*. Verfügbar unter: http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionD.pdf, abgerufen am 05.08.2012.

Beacco J.-C. und Porquier, R. (2007): *Niveau A1 pour le français. Un référentiel.* Paris: Editions Didier.

Beacco J.-C. und Porquier, R. (2008): *Niveau A2 pour le français. Un référentiel.* Paris: Editions Didier.

Beacco, J.-C.; Bouquet, S. und Porquier, R. (2004): *Niveau B2 pour le français. Un référentiel.* Paris: Editions Didier.

Bolton, S.; Glaboniat, M.; Lorenz, H.; Perlmann-Balme, M. und Steiner, S. (2008): *Mündlich. Mündliche Produktion und Interaktion Deutsch: Illustration der Niveaustufen des Gemeinsamen europäischen Referenzrahmens.* München: Langenscheidt.

Bond, T. G. und Fox, C. M. (2007): *Applying the Rasch model: fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum.

Brennan, R. L. (1992): *Generalizability Theory, Instructional Topics in Educational Measurement Series 14.* Abrufbar unter *http://ncme.org/publications/items/*, abgerufen am 05.08.2012.

Briggs, D. C.; Haertel, E.; Schilling, S. G.; Marcoulides, G. A. und Mislevy, R. J. (2004): *Comment: Making an Argument for Design Validity Before Interpretive Validity.* In: Measurement: Interdisciplinary Research & Perspective 2 (3), 171–191.

Camilli, G. und Shepard, L. A. (1994): *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage.

Canale, M. und Swain, M. (1981): *A theoretical framework for communicative competence.* In: Palmer, A. S.; Groot, P. J. und Trosper, S. A. (Hrsg.): *The Construct Validation of Tests of Communicative Competence.* Washington DC: TESOL.

Carr, N. T. (2008): *Using Microsoft Excel® to Calculate Descriptive Statistics and Create Graphs.* In: Language Assessment Quarterly 5 (1), 43.

CEFTrain (2005): CEFTrain. Webseite. http://helsinki.fi/project/ceftrain/index.html, abgerufen am 05.08.2012.

Chapelle, C. A.; Enright, M. K. und Jamieson, J. M. (2007): *Building a Validity argument for the Test of English as a Foreign Language.* Oxford: Routledge.

CIEP (2009): *Productions orales illustrant les 6 niveaux du Cadre européen commun de référence pour les langues.* Webseite. www.ciep.fr/publi_evalcert, abgerufen am 05.08.2012.

CIEP/Eurocentres (2005): *Exemples de productions orales illustrant, pour le français, les niveaux du Cadre européen commun de* référence *pour les langues.* DVD. Strasbourg: Council of Europe.

Cizek, G. J. (1996): Standard-setting guidelines, Instructional Topics in Educational Measurement Series. Abrufbar unter http://ncme.org/publications/items/, abgerufen am 05.08.2012.

Cizek, G. J. und Bunch, M. B. (2006): *Standard Setting: A Guide To Establishing And Evaluating Performance Standards On Tests.* Thousand Oaks, CA: Sage.

Cizek, G. J.; Bunch, M. B. und Koons, H. (2004): *Setting performance standards: contemporary methods, Instructional Topics in Educational Measurement Series.* http://ncme.org/linkservid/8188D217–1 320-5CAE-6EA9C0FC1232764F/showMeta/0/, abgerufen am 05.08.2012.

Clauser, B. E. und Mazor, K. M. (1998): *Using statistical procedures to identify differentially functioning test items, Instructional Topics in Educational Measurement Series.* http://ncme.org/linkservid/80E850A0–1 320-5CAE-6E0F967D1CCD586B/showMeta/0/, abgerufen am 05.08.2012.

Cook, L. L. und Eignor, D. R. (1991): *IRT Equating Methods.* Instructional Topics in Educational Measurement Series 10. http://ncme.org/linkservid/6613B66C-1 320–5CAE-6EA0B4439E8ACC08/showMeta/0/, abgerufen am 05.08.2012.

Corrigan, M. (2007): *Seminar to calibrate examples of spoken performance, Universit*à *per Stranieri di Perugia, CVCL (Centro per la Valutazione e la Certificazione Linguistica) Perugia, 17th–18th December 2005.* http://www.coe.int/T/DG 4/Portfolio/documents/Report_Seminar_Perugiao5.pdf, abgerufen am 03.07.2010.

Coste, D. (2007): *Contextualising Uses of the Common European Framework of Reference for Languages.* Paper presented at Council of Europe Policy Forum on use of the CEFR, Strasbourg 2007. www.coe.int/t/dg4/linguistic/Source/SourceForum07/D-Coste_Contextualise_EN.doc, abgerufen am 05.08.2012.

Council of Europe (1996): *Users' Guide for Examiners.* Strasburg: Language Policy Division.

Council of Europe (1998): *Modern Languages: learning, teaching, assessment. A Common European Framework of Reference.* Strasburg: Language Policy Division.

Council of Europe (2001): *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe (2004a): *Common European Framework of Reference for Languages: Learning, teaching, assessment.* http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf, abgerufen am 05.08.2012.

Council of Europe (2005): *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR).* Reading and Listening Items and Tasks: Pilot Samples illustrating the common reference levels in English, French, German, Italian and Spanish, CD, Strasburg: Council of Europe.

Council of Europe (2006a): *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment.* Writing samples. http://www.coe.int/T/DG4/Portfolio/documents/exampleswriting.pdf, abgerufen am 05.08.2012.

Council of Europe (2006b): *TestDaF Sample Test Tasks.* http://www.coe.int/T/DG4/Portfolio/documents/ALTECEFR%20Writing%20Grid-2.0_TestDaF%20samples.pdf, abgerufen am 05.08.2012.

Council of Europe (2009): *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) – A manual.* http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf, abgerufen am 05.08.2012.

Council of Europe und CIEP (2009): *Productions orales illustrant les 6 niveaux du Cadre européen commun de référence pour les langues.* DVD. Strasbourg und Sèvres: Council of Europe and CIEP.

Davidson, F. und Lynch, B. K. (2002): *Testcraft: A teacher's guide to writing and using language test specifications.* New Haven: Yale University Press.

Davies, A. (Gast-Hrsg.) (1997): Ethics in language testing. Language Testing 14 (3).

Davies, A. (Gast-Hrsg.) (2004): Language Assessment Quarterly 2 & 3.

Davies, A. (2010): *Test fairness: a response.* In: Language Testing 27 (2),171–176.

Davies, A.; Brown, A.; Elder, C.; Hill, K.; Lumley, T. und McNamara, T. (1999): *Dictionary of language testing* (Studies in Language Testing volume 7). Cambridge: Cambridge University Press.

Downing, S. M. und Haladyna, T. M. (2006): *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum.

EALTA (2006): *EALTA Guidelines for Good Practice in Language Testing and Assessment.* http://www.ealta.eu.org/documents/archive/guidelines/English.pdf, abgerufen am 05.08.2012.

Eckes, T. (2009): *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section H:* Many-Facet Rasch Measurement. http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf, abgerufen am 05.08.2012.

Education Testing Services (2002): *ETS Standards for Quality and Fairness.* Princeton, NJ: ETS.

Embretson, S. E. (2007): *Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure?* In: Educational Researcher 36 (8), 449.

Eurocentres und Federation of Migros Cooperatives (2004): *Samples of oral production illustrating, for English, the levels of the Common European Framework of Reference for Languages.* DVD. Strasburg: Council of Europe.

Europarat; Council for Cultural Co-operation, Education Committee, Modern Languages Division; Goethe-Institut Inter Nationes u.a. (Hg.) (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen.* Berlin, München: Langenscheidt.

Figueras, N.; Kuijper, H.; Tardieu, C.; Nold, G. und Takala, S. (2005): *The Dutch Grid Reading/Listening.* Webseite. Zugriffsdatum: 04 / 03/09. Verfügbar auf: http://www.lancs.ac.uk/fss/projects/grid/

Figueras, N.; Noijons, J. (Hrsg.) (2009): *Linking to the CEFR levels: Research perspectives. CITO/Council of Europe.* http://www.coe.int/t/dg4/linguistic/Proceedings_CITO_EN.pdf, abgerufen am 05.08.2012.

Frisbie, D. A. (1988): *Reliability of Scores From Teacher-Made Tests.* Instructional Topics in Educational Measurement Series 3. Abrufbar unter http://ncme.org/publications/items/, abgerufen am 05.08.2012.

Fulcher, G. und Davidson, F. (2007): *Language Testing and Assessment — an advanced resource book.* Abingdon: Routledge.

Fulcher, G. und Davidson, F. (2009): Test architecture, test retrofit, Language Testing 26 (1), 123–144.

GER: siehe Europarat

Glaboniat, M.; Müller, M.; Rusch, P.; Schmitz, H. und Wertenschlag, L. (2005): *Profile Deutsch – Gemeinsamer europäischer Referenzrahmen. Lernzielbestimmungen, Kannbeschreibungen, Kommunikative Mittel, Niveau A1–A2, B1–B2, C1–C2.* Berlin, München: Langenscheidt.

Gorin, J. S. (2007): *Reconsidering Issues in Validity Theory.* In: Educational Researcher 36 (8), 456.

Grego Bolli, G. (Hrsg.) (2008): *Esempi di Produzioni Orali – A illustrazione per l'italiano dei livelli del Quadro comune europeo di riferimento perle lingue*. DVD. Perugia: Guerra Edizioni.

Haertel, E. H. (1999): *Validity arguments for High-Stakes Testing: in search of the evidence*. In: Educational Measurement: Issues and Practice 18 (4), 5.

Hambleton, R. K. und Jones, R. W. (1993): *Comparison of classical test theory and item response theory and their applications to test development, Instructional Topics in Educational Measurement Series 16.* http://ncme.org/linkservid/66968 080–1 320-5CAE-6E4E546A2E4FA9E1/showMeta/0/, abgerufen am 05.08.2012.

Harvill L. M. (1991): *Standard error of measurement.* Instructional Topics in Educational Measurement Series 9. http://ncme.org/linkservid/6606715E-1 320–5CAE-6E9DDC581EE47F88/showMeta/0/, abgerufen am 05.08.2012.

Heaton, J. B. (1990): *Classroom Testing.* Harlow: Longman.

Holland, P. W. und Dorans, N. J. (2006): *Linking and Equating*. In: Brennan, R. L. (Hrsg.) *Educational measurement.* 4th edition, Washington, DC: American Council on Education/Praeger.

Hughes, A. (1989): *Testing for Language Teachers*. Cambridge: Cambridge University Press.

ILTA (2000): *ILTA Code of Ethics.* Abrufbar unter: http://www.iltaonline.com/index.php?option=com_content&task=view&id=57&Itemid=47, abgerufen am 05.08.2012.

ILTA (2007): *ILTA Guidelines for Practice.* Abrufbar unter: http://iltaonline.com/index.php?option=com_content&task=view&id=122&Itemid=i33, abgerufen am 05.08.2012.

Instituto Cervantes (2007): *Plan curricular del Instituto Cervantes – Niveles de referencia para el español.* Madrid: Edelsa.

JCTP (1988): *Code of Fair Testing Practices in Education.* Abrufbar unter: http://www.apa.org/science/programs/testing/fair-code.aspx, abgerufen am 05.08.2012.

JLTA (keine Angabe): *Code of Good Testing Practice.* http://www.avis.ne.jp/—youichi/CP.html, abgerufen am 08.12.2009.

Jones, N. und Saville, N. (2009): *European Language Policy: Assessment, Learning and the CEFR*. In: Annual Review of Applied Linguistics 29, 51–63.

Jones, P.; Smith, R. W. und Talley, D. (2006): *Developing Test Forms for Small-Scale Achievement Testing Systems.* In Downing, S. M. und Haladyna, T. M. (Hrsg.): Handbook of Test Development, Mahwah, NJ: Lawrence Erlbaum.

Jones, R. L. und Tschirner, E. (2006): *A Frequency Dictionary of German – Core Vocabulary for Learners.* New York: Routledge.

Kaftandjieva, F. (2004): *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section B: Standard Setting.* http://www.coe.int/t/dg4/linguistic/CEF-refSupp-SectionB.pdf, abgerufen am 05.08.2012.

Kane, M. (2002): *Validating High Stakes Testing Programs.* In: Educational Measurement: Issues and Practices 21 (1), 31–41.

Kane, M. (2004): *Certification Testing as an Illustration of Argument-Based Validation*. In: Measurement: Interdisciplinary Research & Perspective 2 (3), 135–170.

Kane, M. (2006): *Validation.* In: Brennan, R. L. (Hrsg.): Educational measurement. 4th edition, Washington, DC: American Council on Education/Praeger.

Kane, M. (2010): Validity and fairness, Language Testing 27(2), 177–182.

Kane, M.; Crooks, T. und Cohen, A. (1999): *Validating measures of performance*. In: Educational Measurement: Issues and Practice 18 (2), 5–17.

Kolen, M. J. (1988): *Traditional Equating Methodology.* Instructional Topics in Educational Measurement Series 6. http://www.ncme.org/pubs/items/ii.pdf, abgerufen am 03.05.2009.

Kolen, M. J. (2006): *Scaling and Norming.* In: Brennan, R. L. (Hrsg.): Educational measurement. 4th edition, Washington, DC: American Council on Education/Praeger.

Kuijper, H. (2003): *QMS as a Continuous Process of Self-Evaluation and Quality Improvement for Testing Bodies.* http://www.alte.org/qa/index.php, abgerufen am 07.12.2009.

Kunnan, A. J. (2000a): *Fairness and justice for all*. In: Kunnan, A. J. (Hrsg.): Fairness and validation in language assessment. Cambridge: Cambridge University Press, 1–13.

Kunnan, A. J. (2000b): *Fairness and Validation in Language Assessment: Selected papers from the 79th Language Testing Research Colloquium, Orlando, Florida* (Studies in Language Testing volume 9). Cambridge: Cambridge University Press.

Kunnan, A. J. (2004): *Test Fairness*. In: Milanovic, M. und Weir, C. (Hrsg.): European Language Testing in a Global Context – Proceedings of the ALTE Barcelona Conference, July 2007 (Studies in Language Testing volume 18), Cambridge: Cambridge University Press.

Linacre, J. M. (2009): *Facets 3.64.0.* Software. Winsteps.com software.

Lissitz, R. W. und Samuelsen, K. (2007a): *A Suggested Change in Terminology and Emphasis Regarding Validity and Education.* In: Educational Researcher 36 (8), 437.

Lissitz, R. W. und Samuelsen, K. (2007b): *Further Clarification Regarding Validity and Education.* In: Educational Researcher 36 (8), 482.

Livingston, S. (2004): *Equating Test Scores (Without IRT).* http://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf, abgerufen am 05.08.2012.

McNamara, T. und Roever, C. (2006): *Fairness Reviews and Codes of Ethics*. In: Language Learning 56 (S2), 129–148.

Messick, S. (1989): *Meaning and values in test validation: the science and ethics of assessment.* In: Educational Researcher: Issues and Practice 18, 5–11.

Messick, S. (1989): *Validity*. In: Linn, R. (Hrsg.): Educational measurement. 3rd edition, New York: Macmillan, 13–103.

Mislevy, R. J. (2007): *Validity by Design*. In: Educational Researcher 36 (8), 463.

Mislevy, R. J.; Steinberg, L. S. und Almond, R. G. (2003): *Focus Article: On the Structure of Educational Assessments*. In: Measurement: Interdisciplinary Research & Perspective 1 (1), 3–62.

Moss, P. A. (2007): *Reconstructing Validity.* In: Educational Researcher 36 (8), 470.

North, B. und Jones, N. (2009): *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) — Further Material on Maintaining Standards across Languages, Contexts and Administrations by exploiting Teacher Judgment and IRT Scaling.* http://www.coe.int/t/dg4/linguistic/Manual%20 – %20Extra%20Material%20 – %20proofread%20 – %20FINAL.pdf, abgerufen am 05.08.2012.

Parkes, J. (2007): *Reliability as Argument*. In: Educational Measurement: Issues and Practice 26(4): 2–10.

Perlmann-Balme, M. und Kiefer, P. (2004): *Start Deutsch. Deutschprüfungen für Erwachsene. Prüfungsziele, Testspezifikation.* München, Frankfurt: Goethe-Institut und WBT.

Perlmann-Balme, M.; Plassmann, S. und Zeidler, B. (2009): *Deutsch-Test* für *Zuwanderer. Prüfungsziele, Testspezifikation.* Berlin: Cornelsen.

Saville, N. (2005): *Setting and monitoring professional standards: A QMS approach.* In: Research Notes 22. http://www.cambridgeesol.org/rs_notes/rs_nts22.pdf, abgerufen am 05.08.2012.

Sireci, S. G. (2007): *On Validity Theory and Test Validation*. In: Educational Researcher 36 (8), 477.

Spinelli, B. und Parizzi, F. (2010): *Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1 e B2*. Milan: RCS libri – Divisione education.

Spolsky, B. (1981): *Some ethical questions about language testing*. In: Klein-Braley, C. und Stevenson, D. (Hrsg.): Practice and problems in language testing, Frankfurt: Verlag Peter Lang, 5–21.

Stiggins, R. J. (1987): *Design and Development of Performance Assessment.* Instructional Topics in Educational Measurement Series 1.http://ncme.org/linkservid/3E3AC538–1 320-5CAE-6E780DD57D5D8EE2/showMeta/0/, abgerufen am 05.08.2012.

Traub, R. E. und Rowley, G. L. (1991): *Understanding reliability.* Instructional Topics in Educational Measurement Series 8. http://ncme.org/linkservid/65F3B451–1 320-5CAE-6E5A1C4257CFDA23/showMeta/0/, abgerufen am 05.08.2012.

Trim, J. L. M. (2010): Plenary presentation at ACTFL-CEFR Alignment Conference, Leipzig, Juni 2010.

University of Cambridge ESOL Examinations (2004): *Samples of oral production illustrating, for English, the levels of the Common European Framework of Reference for Language.* DVD. Strasbourg: Council of Europe.

University of Cambridge ESOL Examinations/Council of Europe (2009a): *Common European Framework of Reference for Languages Examples of Speaking Test Performance at Levels A2 to C2*. DVD. Cambridge: University of Cambridge ESOL Examinations.

University of Cambridge ESOL Examinations/Council of Europe (2009b): *Common European Framework of Reference for Languages Examples of Speaking Test Performance at Levels A2 to C2.* Webseite. http://

www.cambridgeesol.org/what-we-do/research/speakridgeesol.org/what-we-do/research/speaking-performances.html, abgerufen am 09.02.2009.

van Avermaet, P. (2003): *QMS and The Setting of Minimum Standards: Issues of Contextualisation Variation between The Testing Bodies.* http://www.alte.org/q a/index.php, abgerufen am 07.12.2009.

van Avermaet, P.; Kuijper, H. und Saville, N. (2004): *A Code of Practice and Quality Management System for International Language Examinations.* In: Language Assessment Quarterly 1 (2 & 3), 137–150.

van Ek, J. A. und Trim, J. (1990): Waystage 1990, Cambridge: Cambridge University Press.

van Ek, J. A. und Trim, J. (1991): Threshold 1990, Cambridge: Cambridge University Press.

van Ek, J. A. und Trim, J. (2001): Vantage, Cambridge: Cambridge University Press.

Verhelst, N. (2004a): *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section C: Classical Test Theory.* http://www.coe.int/t/dg4/linguistic/CEF-refSupp-SectionC.pdf, abgerufen am 05.08.2012.

Verhelst, N. (2004b): *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section E: Generalizability Theory.* http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionE.pdf, abgerufen am 05.08.2012.

Verhelst, N. (2004c): *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section F: Factor Analysis.* http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionF.pdf, abgerufen am 05.08.2012.

Verhelst, N. (2004d): *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF: section G: Item Response Theory.* http://www.coe.int/t/dg4/linguistic/CEF-ref-supp-SectionG.pdf, abgerufen am 05.08.2012.

Ward, A. W. und Murray-Ward, M. (1994): *Guidelines for Development of item banks, Instructional Topics in Educational Measurement Series 17.* http://ncme.org/linkservid/6A9EABFE-1 320–5CAE-6EC8622570FA7AA3/showMeta/0/, abgerufen am 05.08.2012.

Weir, C. J. (2005): *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Weir, C. J. und Milanovic, M. (Hrsg.) (2003): *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002* (Studies in Language Testing volume 15). Cambridge: Cambridge University Press.

Widdowson, H. G. (1978): *Language Teaching as Communication*. Oxford: Oxford University Press.

Xi, X. (2010): *How do we go about investigating test fairness?* In: Language Testing 27(2): 147–170.

Webseiten (alle abgerufen am 05.08.2012):

Association of Language Testers in Europe: www.alte.org

English Profile: www.englishprofile.org

European Association for Language Testing and Assessment: http://www.ealta.eu.org/

International Language Testing Association: http://www.iltaonline.com/

Language Policy Division, Council of Europe: http://www.coe.int/t/dg4/linguistic/default_EN.asp

ANHÄNGE

Anhang I: Aufbau einer Beweisführung zur Validität

Dieser Anhang stellt einen Validatitätsansatz vor, der auf der Ausarbeitung einer Beweisführung oder Argumentation zur **Validität** beruht. Er ist umfangreicher als die in Kapitel 1.2.3 aufgeführten Grundzüge und zeigt, dass die einzelnen Schritte in der Beweis – bzw. Argumentationskette nicht als isoliert und starr fortlaufend zu betrachten sind, sondern sich vielmehr überlappen und in engem Bezug zueinander stehen.

Bei Kane (2006), Kane, Crooks und Cohen (1999), Bachman (2005) und Bachman und Palmer (2010) finden sich ausführlichere Hinweise zum Beleg von Validität. Validierung ist demnach ein stetiger Prozess, der mit der Zeit immer mehr und immer genauer ausgeführte Belege der Validität aufführt.

Im Zentrum der Beweisführung zur Validität stehen die Interpretation und die Verwendung von Testergebnissen. Damit folgt man der Definition von Validität als Ausmaß, in dem theoretisch und empirisch begründete Schlussfolgerungen die Interpretation von Testergebnissen gemäß der intendieren Verwendung des Tests untermauern (AERA et al 1999).

Eine Validitätsargumentation besteht also aus einer Reihe von Behauptungen, die beschreiben, warum die empfohlenen Interpretationen der Testergebnisse valide sind, und die dies entsprechend belegen. Dieser Anhang gibt einen Überblick über den Aufbau einer solchen Beweisführung.

Die Präsentation der Argumentation gegenüber den **Beteiligten** beginnt mit der klaren Aussage, wie Testergebnisse für einen bestimmten Zweck interpretiert werden sollen. Die Beweisführung hinsichtlich der **Verwendung des Tests** erklärt diese Aussage. Das, was wir Validitätsbeleg nennen, ist also im Grunde genommen die empirisch und theoretisch gestützte Verwendungsbegründung.

Abbildung 16 zeigt die konzeptionelle Sicht einer begründenden Argumentation nach Bachman (2005). Es handelt sich um eine logische Folge, bestehend aus vier Schritten (jeder durch einen Pfeil dargestellt), die die Verwendung der Testergebnisse rechtfertigt. Jeder Schritt bietet die konzeptionelle Grundlage für den nächsten. So sind z. B. allgemeine Testergebnisse (*universe score*) nur dann sinnvoll, wenn sie die im Test beobachtete Leistung (*observed score*) angemessen zeigen. Das Diagramm zeigt keine Abfolge von Phasen, in der eine nach der anderen abgeschlossen sein muss. Belege für jeden einzelnen Schritt können auch aus anderen Phasen der Testentwicklung und Versionsgenerierung gewonnen werden.

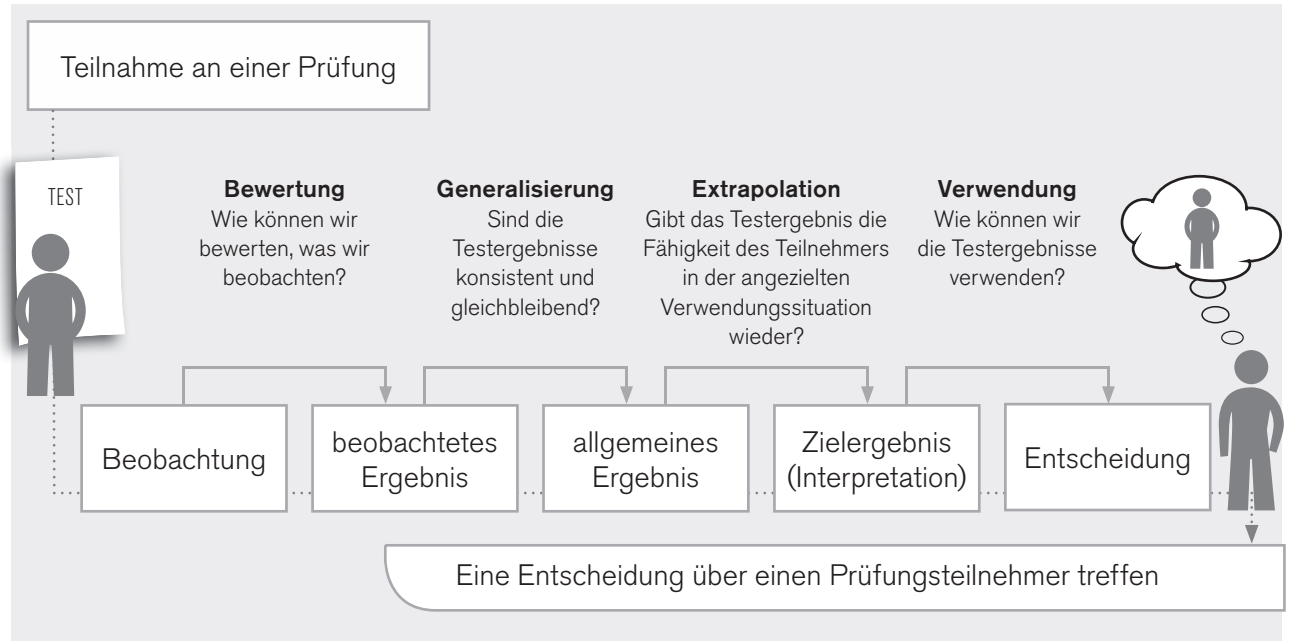


Abb. 16: Logische Folge in einer Validitätsargumentation (angepasst von Cane, Crooks, Cohen 1999, Bachman 2005)

Der Validitätsbeleg stützt die Verwendungsargumentation und besteht aus empirisch wie theoretisch begründeten Nachweisen sowie Belegen aus dem praktischen Verwendungszusammenhang. Belege für jeden einzelnen Schritt werden während der Phasen der Testentwicklung, der Versionsgenerierung und des Testeinsatzes gewonnen.

Viele Belege für die Validitätsargumentation werden aus dem routinemäßigen Einsatz des Tests kommen. Beispiele hierfür sind in Kapitel 6.1 aufgeführt. Solche Belege werden für einen anderen, unmittelbaren Zweck gesammelt, etwa für die Überprüfung der Bewerterleistungen, sind aber auch für den Aufbau einer Validitätsargumentation von Nutzen. Dies wird in Abbildung 17 dargestellt.

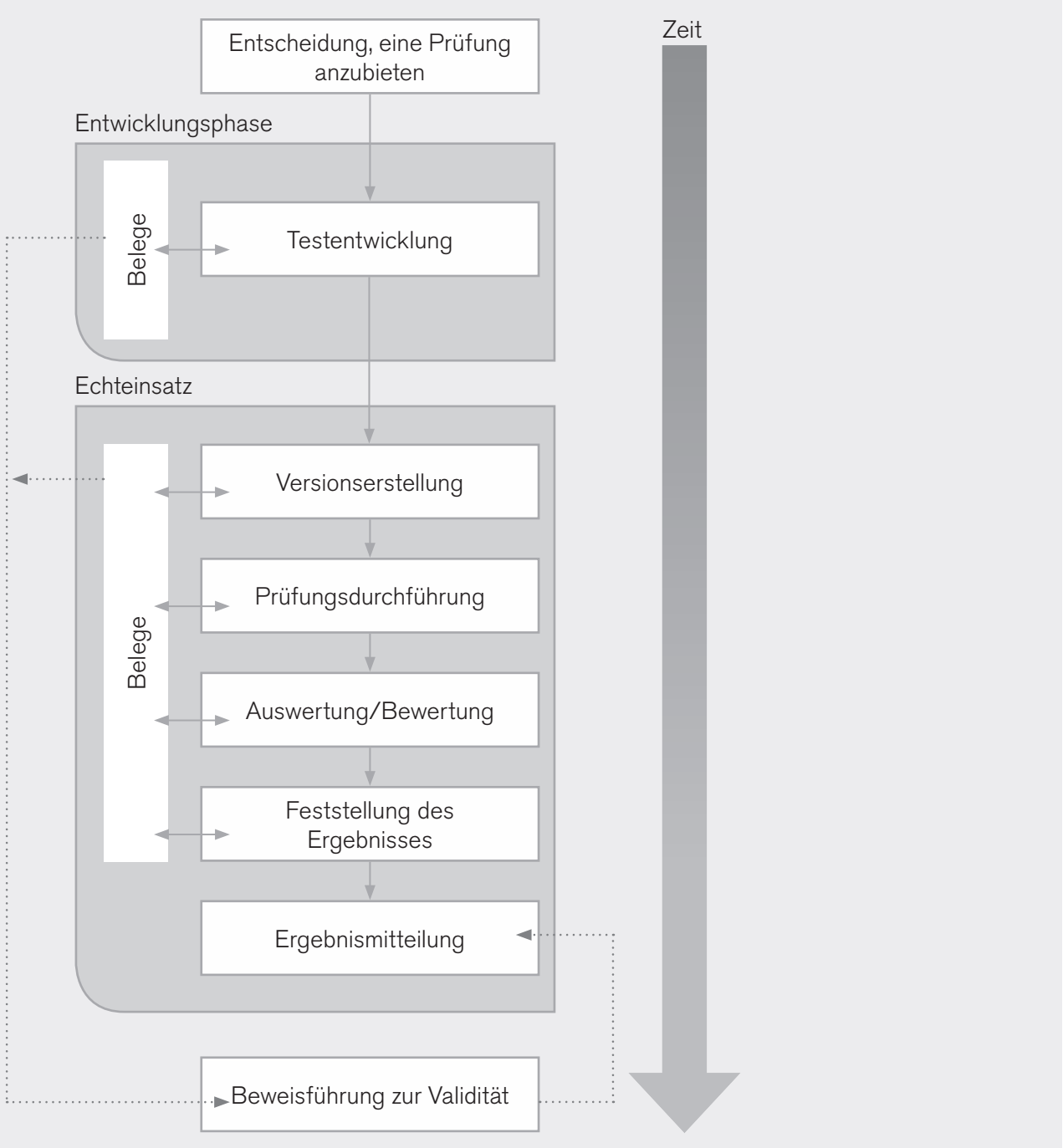


Abb. 17: Testentwicklungszyklus, periodische Überprüfung und Validitätsbelege

Die Beweisführung zur Validität kann weiterentwickelt und verbessert werden, indem bei jeder Generierung einer neuen Testversion und deren Einsatz diese Belege genutzt werden. Der Aufbau eines Validitätsnachweises sollte in einer sehr frühen Phase des Prozesses beginnen, nämlich wenn der Zweck des Tests definiert wird. Es kann zudem vieles aus der Validitätsargumentation für ein Testformat auch für ein weiteres genutzt werden.

Einige Wissenschaftler (Bachman 2005, Mislevy et al 2003) betonen, dass eine Beweisführung zur Validität die Form einer informellen Argumentation haben sollte, nicht die einer formell logischen Argumentation. Das heißt, dass die Beweisführung nicht allein durch logische Schlussfolgerungen als richtig oder falsch nachgewiesen werden kann. Vielmehr kann sie bei einer Überprüfung als mehr oder weniger überzeugend eingeschätzt werden. Wie überzeugend sie ist, hängt von den erbrachten unterstützenden Belegen ab.

Neue oder neu interpretierte Belege sowie neue wissenschaftliche Erkenntnisse können dazu führen, dass die ursprüngliche Beweisführung zur Validität weniger überzeugend erscheint. Auch kann es vorkommen, dass Testanbieter unabsichtlich ihre bevorzugte Interpretation einfach immer wieder bestätigen, ohne kritisch genug zu sein. Nachdem die Beweisführung zum ersten Mal aufgestellt worden ist, muss sie daher immer wieder hinterfragt werden, auch wenn dies eine Änderungen der empfohlenen Interpretation der Testergebnisse zur Folge haben kann. Dies kann beispielsweise geschehen, indem man nach anderen Auslegungen der Belege sucht oder überprüft, ob alle Rückschlüsse in der Argumentation tatsächlich stimmig sind. Der Testanbieter muss also seine Beweisführung überprüfen, ggf. Änderungen vornehmen und Gründe dafür angeben, warum die Belege in bestimmter Weise interpretiert wurden.

Beispiele für verschiedene Belege zur Unterstützung einer Beweisführung zur Validität finden sich auf den folgenden Seiten dieses Anhangs. Zusätzlich werden Beispiele für verschiedene Interpretationsmöglichkeiten aufgezeigt. Alle Beispiele basieren auf der Arbeit von Kane (2004) und Bachman (2005). Sie werden gemäß der Struktur dieses Handbuchs präsentiert: Testentwicklung, Generierung von Echttestversionen, Prüfungsdurchführung, Auswertung, Bewertung und Benotung und Übermittlung der Testergebnisse. Testanbieter können diese Belege als Ausgangspunkt für die eigene Arbeit an einem Validitätsnachweis nutzen. Die Listen sind jedoch nicht erschöpfend.

Weiterführende Literatur

ALTE (2005: 19) gibt eine nützliche Zusammenfassung der Formen von Validität und zeigt die Hintergründe des heutigen Validitätskonzepts.

AERA et al. (1999) geben einen Überblick über die zeitgemäße Auffassung von Validität und Standards, betont spezielle wichtige Punkte und kann somit für den Aufbau eines Validitätsnachweises hilfreich sein.

Messick (1989) spricht vom unitären Validitätskonzept und betrachtet auch dessen ethischen Aspekte.

Haertel (1999) gibt ein Beispiel zur Verbindung zwischen Validitätsbelegen und Argumentation und der Interpretation der Testergebnisse.

Kane, Crooks und Cohen (1999) zeichnen ein klares Bild der ersten Phasen der Validitätsargumentation. Detailliertere Abhandlungen finden sich bei Kane (2006).

Bachman (2005) analysiert die Validitätsargumentation im Hinblick auf Sprachprüfungen und stellt das von Bachman und Palmer (1996) erstellte Model zusammen mit der Validitätsargumentation bildlich dar. Im früheren Model wurde die Zweckmäßigkeit, also der Ausgleich zwischen Reliabilität, Validität, Authentizität, Interaktivität und Wirkung als der wichtigste Aspekt eines Tests angesehen.

Bachman und Palmer (2010) zeigen die zentrale Bedeutung der Validitätsargumentation bei der Testentwicklung und setzen hierfür eine Grundstruktur.

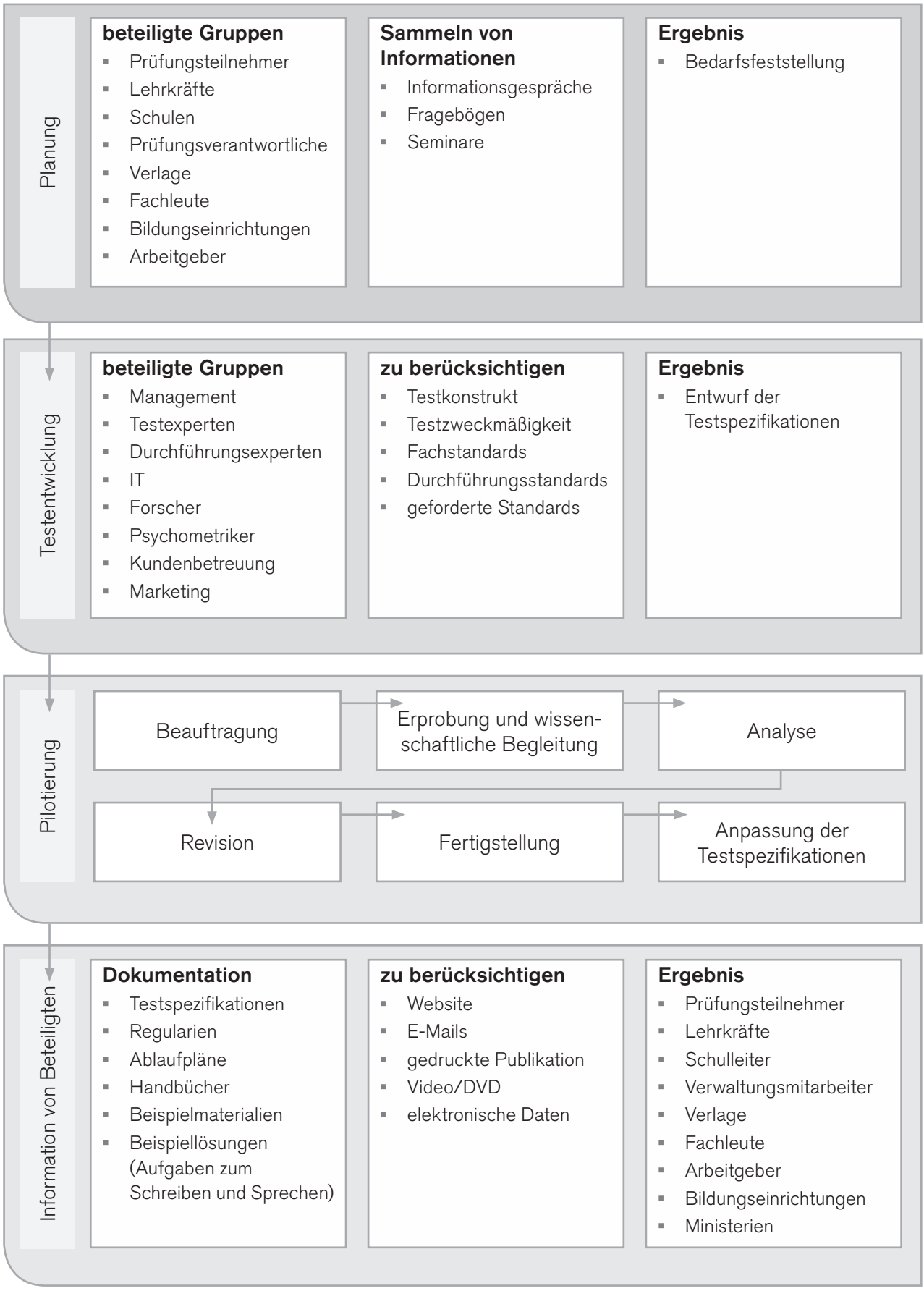
	Bewertung	Verallgemeinerung	Übertragung	Verwendung
	Wie können wir bewerten, was wir beobachten?	Sind die Ergebnisse gleichbleibend und beständig?	Entspricht das Testergebnis den Fähigkeiten des Teilnehmers in der angestrebten Sprachverwendungssituation?	Wie können wir das Testergebnis verwenden?
Testentwicklung (Kapitel 2)				
Belege dafür		Die Testspezifikationen verlangen ein bestimmtes Testformat – dies stützt die Annahme einer Übereinstimmung zwischen den verschiedenen Testversionen (siehe Kapitel 2 und Anhang III).	Der Leitfaden für Testautoren sowie die Testspezifikationen definieren eine Verwendungssituation. Diese kann auch durch eine Bedarfsanalyse identifiziert werden (siehe Kapitel 2.4). Belege für die angemessene Festsetzung der Bestehensgrenze unterstützen die empfohlene Interpretation der jeweiligen Testergebnisse (siehe Kapitel 2.0 und 5.2).	
Belege dagegen			Einige Bereiche des Konstrukts werden nicht vollständig in den Testspezifikationen aufgezeigt. Dies würde bedeuten, dass die Testergebnisse keine angemessenen Informationen darüber geben, was der Teilnehmer tatsächlich kann (siehe Kapitel 1.1 und 2).	

	Bewertung	Verallgemeinerung	Übertragung	Verwendung
Generierung von Testversionen (Kapitel 3)				
Belege dafür	Alle Lösungsschlüssel sind korrekt. Grammatikbücher, Wörterbücher und Fachwissen können dies belegen.	Die Items einer Testversion bilden das Konstrukt genauso ab wie die Items einer anderen Testversion. Es ist unwahrscheinlich, dass jedes Mal genau der gleiche Bereich abgedeckt wird, aber die Konstruktbereiche müssen so gewählt werden, dass sie vergleichbar sind (siehe Kapitel 2, 3, 5 und Anhang VII). Die Verlinkung der Versionen ist angemessen (siehe Anhang VII). Bei der Verwendung von statistischen Analysen wurden geringe Messabweichungen gefunden, und die statistischen Modelle wurden an die Werte angepasst (siehe Anhang VII).	Fachleute wurden zur Item- und Testerstellung herangezogen (siehe Kapitel 3.2).	
Belege dagegen		Testversionen sind nicht miteinander abgeglichen. Testversionen beziehen sich nicht auf das gleiche Konstrukt.	Einige Bereiche des Konstrukts werden ggf. im Testmaterial nicht angemessen abgebildet. So würden die Ergebnisse nicht die gewünschten Informationen dazu liefern, was der Teilnehmer kann.	

	Bewertung	Verallgemeinerung	Übertragung	Verwendung
Prüfungsdurchführung (Kapitel 4)				
Belege dafür	In der Prüfungsdurchführung werden alle Regeln befolgt. Das hilft, um zu zeigen, dass das Testergebnis nicht von anderen Faktoren (wie zu wenig oder zu viel Zeit) beeinflusst wird (siehe Kapitel 4.2).	Die Prüfungsordnung wurde stets befolgt. Dies unterstützt den Nachweis, dass die Durchführung bei Einsatz verschiedener Testversionen gleichbleibend ist (siehe Kapitel 4.2).		
Belege dagegen	Ungeahndete Täuschungsversuche führen dazu, dass die Testergebnisse nicht die Fähigkeiten des Teilnehmenden widerspiegeln.	Unbemerkte Täuschungsversuche führen dazu, dass die Ergebnisse einiger Teilnehmer nicht angemessen ihre Sprachfähigkeiten widerspiegeln. Dies kann Unterschiede zwischen verschiedenen Testversionen hervorrufen.	Äußere Faktoren hatten ggf. einen Einfluss auf die Testergebnisse. Dies kann daran liegen, dass die Prüfungsordnung nicht befolgt wurde. So spiegeln auch die Testergebnisse die äußeren Faktoren wider und man kann nicht sicher sagen, dass sie die Sprachfähigkeit korrekt anzeigen (siehe Kapitel 4.1).	

	Bewertung	Verallgemeinerung	Übertragung	Verwendung
Auswertung, Benotung und Übermittlung der Ergebnisse (Kapitel 5)				
Belege dafür	Beim Auswerten wurden standardisierte Verfahren eingehalten. Dies unterstützt den Nachweis, dass die Testergebnisse nicht von anderen Faktoren (z. B. falsche Lösungsschlüssel, Fehler beim Scannen) beeinflusst wurden (siehe Kapitel 5.0). Die Bewertung war korrekt und reliabel (siehe Kapitel 5.1 und Anhang VIII).	Belege für die Reliabilität der Testergebnisse (normalerweise statistische Belege) können zeigen, dass eine Testversion die Teilnehmenden in zuverlässig bewertet (siehe Kapitel 1.3, 5.1 und Anhang VII). Wenn nur die Daten von wenigen Teilnehmenden analysiert wurden, so sind diese doch für die gesamte Teilnehmergruppe repräsentativ (siehe Anhang VII). Bestehensgrenzen mit wenigen Messabweichungen bedeuten, dass die Teilnehmenden mit größerer Wahrscheinlichkeit auf der richtigen Seite der Punktegrenze platziert werden (siehe Anhang VII).	Bei Bewertung durch Fachleute ist es wahrscheinlich, dass die Bewertungen das zeigen, was gezeigt werden soll. (siehe Kapitel 5.1). In ähnlicher Weise erhöht die Nutzung von gut geschriebenen Bewertungsanleitungen die Chance, dass die Leistung so bewertet wird, wie sie bewertet werden soll (siehe Kapitel 2.5 und 5.1.3).	Wenn Entscheidungen auf der Grundlage der Testergebnisse nach bestimmten Regeln getroffen werden, wird der Test wahrscheinlich wie geplant verwendet und unerwünschte Auswirkungen werden reduziert (siehe Kapitel 1.2, 5.3 und Anhang I).
Belege dagegen		Wenn Daten einer nicht-repräsentativen Gruppe von Teilnehmenden zur Analyse verwendet werden, wird die Analyse ggf. Messabweichungen und / oder Verzerrungen enthalten (siehe Anhang VII).		Wenn bei der Entscheidungsfindung keine Regeln und Standardverfahren befolgt werden, wird der Test ggf. unangemessen verwendet (siehe Kapitel 1.5 und 5.3).

Anhang II: Der Prozess der Testentwicklung



Anhang III: Beispiel für ein Testformat

Deutsches Beispiel

Inhalt und Überblick

Prüfungsteil	Zeit	Aufgabentyp		Itemzahl	Testfokus
Leseverstehen	90 min	Teil 1	Zuordnungsaufgabe Zeitschriftenartikel, kurzer Aufsatz/ Essay o. Ä. 400–500 Wörter 7 Lücken, in jede Lücke muss der passende Satz zur Vervollständigung des Texts eingesetzt werden.	6	Testrekonstruktion Längere Texte verstehen und Lücken schließen
		Teil 2	Zuordnungsaufgabe Sachtext aus Zeitung, Zeitschrift o. Ä. ca. 650–850 Wörter 6 Items, die den Textabschnitten zugeordnet werden müssen.	6	Selektives Verstehen Zentrale Aussagen eines längeren Texts verstehen
		Teil 3	Aufgabe richtig/falsch/nicht im Text Sachtext aus Zeitung, Zeitschrift o. Ä. 1 000–1 200 Wörter 11 Items, bei denen entschieden werden muss, ob die jeweilige Aussage im Hinblick auf die im Text enthaltenen Informationen richtig, falsch oder nicht im Text gegeben ist.	11	Detailverstehen Detailinformationen in längeren Texten verstehen
			Makroaufgabe 1 Item, bei dem unter 3 möglichen Überschriften die passende gewählt werden muss.	1	Globalverstehen Text als Ganzes verstehen
Sprachbausteine		Teil 1	4er-Mehrfachwahlaufgabe Sachtext aus Zeitung, Zeitschrift o. Ä. ca. 320–350 Wörter 22 Items (zu Grammatik, Lexik, Rechtschreibung), bei denen das jeweils passende Wort zu Textlücken zu wählen ist.	22	Grammatik und Rechtschreibung Grammatik- und Rechtschreibkompetenzen unter Beweis stellen

Hör-verstehen	ca. 40 min	Teil 1	Zuordnungsaufgabe Kurze Interviews mit einzelnen Personen ca. 100–140 Wörter 8 Items (und 10 Alternativen zur Zuordnung). 8 Hörtexten müssen passenden Aussagen zugeordnet werden.	8	Globalverstehen Die zentrale Aussage, Intention o. Ä. eines Sprechers, bzw. einer Sprecherin verstehen
		Teil 2	3er-Mehrfachwahlaufgabe Interview ca. 1 100 Wörter 10 Items. Es muss jeweils entschieden werden, welche von 3 Aussagen im Interview gehört wurde.	10	Detailverstehen Aussagen, Haltungen eines Sprechers, bzw. einer Sprecherin in einem Interview verstehen
		Teil 3	Informationen ergänzen Monolog (Vortrag / Vorlesung) ca. 10 Minuten (= ca. 1 100 Wörter) 10 Items. Vortrag wird gehört, anschließend müssen zu Folien, die die wesentlichen Inhalte des Vortrages enthalten, fehlende Informationen stichpunktartig ergänzt werden.	10	Informationstransfer Einem längeren Hörtext (Vortrag, Vorlesung etc.) folgen und zentrale Punkte aus dem Inhalt schriftlich festhalten können
Schriftlicher Ausdruck	70 min		Erörterung, Stellungnahme etc. Inputtext 1+2 als Schreibanlass; die zwei gegensätzlichen Aussagen in der gewählten Aufgabe sind gemeinsam der Schreibanlass. Textlänge: ca. 45–55 Wörter geforderte Textlänge: mindestens 350 Wörter	1 Aufgabe aus 2 Aufgaben zur Auswahl	Text schreiben Einen komplexeren Text (Erörterung, Essay o. Ä.) schreiben können
Mündlicher Ausdruck	16 min		Paarprüfung	1 Thema aus 2 Themen/ Fragestellungen zur Auswahl	
		Teil 1a	Aufgabentext (standardisiert), 2 Themen zur Auswahl Kurzreferat in Form von Monolog ca. 3 Minuten pro Person		Präsentation Fähigkeit, ein Thema gut strukturiert, flüssig und sprachlich angemessen präsentieren zu können.
		Teil 1b	Aufgabentext (standardisiert), 2 Themen zur Auswahl Den Vortrag des Partners bzw. der Partnerin zusammenfassen und Anschlussfragen stellen ca. 2 Minuten pro Person	verschieden	Zusammenfassung / Anschlussfragen Fähigkeit, eine Präsentation in den wichtigsten Punkten zusammenzufassen und Nachfragen zu stellen sowie auf Nachfragen auch angemessen antworten zu können
		Teil 2	Zitat eines Schriftstellers, Wissenschaftlers o. Ä. und standardisierte Fragen bzw. Punkte für die Diskussion ca. 6 Minuten	verschieden	Diskussion Fähigkeit, über ein Thema sprachlich angemessen diskutieren zu können und dabei sowohl sprachliche als auch Kompetenzen der sprachlichen Interaktion einzusetzen.

Beispiel für Leseverstehen

Allgemeine Beschreibung	
Format	Der Subtext „Leseverstehen“ umfasst verschiedene Textsorten, deren gemeinsamer Bezugspunkt jedoch in einer Relevanz für das Studium liegt: Artikel, Berichte aus Zeitungen, Zeitschriften oder von Internetseiten, populärwissenschaftliche Texte oder Essays
Zeit	90 Minuten
Anzahl der Teile	4 Aufgaben (inklusive des separaten Teils „Sprachbausteine“)
Anzahl der Aufgaben	24
Aufgabentypen	6 Zuordnungsaufgaben 6 Zuordnungsaufgaben 11 Aufgaben richtig/falsch/nicht im Text 1 Makroaufgabe
Texttypen	Textsorten, deren gemeinsamer Bezugspunkt in einer Relevanz für das Studium liegt: Artikel, Berichte aus Zeitungen, Zeitschriften oder von Internetseiten, populärwissenschaftliche Texte oder Essays, die eher deskriptiv oder eher argumentativ sein können.
Textlänge	400–500 Wörter 650–850 Wörter 1 000–1 200 Wörter 320–350 Wörter
Antwortformat	Teilnehmende markieren ihre Antworten auf dem Antwortbogen oder schreiben ein Wort auf ein computerlesbares Antwortblatt.
Bewertung	Leseverstehen 1–3: Für jede korrekte Lösung gibt es zwei Punkte Sprachbausteine: Für jede korrekte Lösung gibt es einen Punkt.

Anhang IV: Hinweise für Testautoren

Hinweise zur Textauswahl

Die Definition von „Text“ in diesem Handbuch lehnt sich an die Definition in Abschnitt 4.6 des GER an. Ein Text ist jede Art von sprachlichem Produkt, sei es eine gesprochene Äußerung oder etwas Geschriebenes.

Testautorinnen und -autoren benötigen Hinweise zur Auswahl von Texten. Diese sollten folgende Punkte berücksichtigen:

- die besten Textquellen (z. B. Qualitätszeitungen, Broschüren)
- Quellen, die wahrscheinlich keine guten Texte abgeben (z. B. Fachliteratur)
- ein allgemeiner Hinweis zur Vermeidung von Verzerrungen der Testergebnisse (aufgrund Kultur, Geschlecht, Alter etc.)

Auch können Gründe angegeben werden, warum frühere Texte abgelehnt wurden, zum Beispiel:

- zu viel vorausgesetztes kulturspezifisches oder lokales Wissen (es sei denn, dies soll getestet werden)
- Themen, die für die Zielgruppe als nicht angemessen erachtet werden, z. B. Krieg, Tod, politische oder religiöse Einstellungen oder andere Themen, die einige Prüfungsteilnehmende verletzen oder aufregen können
- Themen außerhalb des Erfahrungsschatzes der teilnehmenden Altersgruppe
- zu schwierige oder zu leichte Vokabeln oder sprachliche Konzepte
- sprachliche oder stilistische Fehler oder Eigenarten
- schlechte redaktionelle Bearbeitung des Originaltexts

Es ist auch möglich, eine Liste von Themen zu geben, die in der Vergangenheit schon ausreichend abgedeckt wurden, so dass sie nicht mehr benötigt werden.

Kapitel 4 und 7 des GER bieten eine gute Hilfestellung beim Suchen von angemessenen Texten und stellen die Textvorschläge in den Kontext des allgemeinen Ansatzes des Europarats zum Erlernen von Sprachen. Die in Abschnitt 4.6.2 aufgeführten Medien (Stimme, Telefon, Radio etc.) zusammen mit den in Abschnitt 4.6.3 beschriebenen gesprochenen und geschriebenen Textarten stellen eine nützliche Checkliste dar und bieten die Möglichkeit, Aufgabentypen zu variieren.

Formale Hinweise

Man kann Testautorinnen und -autoren folgende formale Hinweise geben:

- Zeilenabstand bei getippten Texten
- Informationen in der Kopfzeile
- Einreichung der Originaltexte (ggf. als Fotokopie)
- Notwendige Angaben zu den Textquellen (z. B. Veröffentlichungsdatum)

Detaillierte Hinweise zu jeder Aufgabe

Dies wird am besten anhand eines fiktiven Beispiels dargestellt: Für einen modifizierten C-Test, der sich mehr auf Strukturwörter als auf allgemeine Lexik bezieht, erhält der Testautor folgende Hinweise:

- Ein authentischer Text von ca. 200 Wörtern wird benötigt. Er sollte eine kurze Überschrift haben. Der Schwerpunkt liegt auf einzelnen Strukturwörtern. Es sollten nicht zu viele unbekannte Vokabeln vorkommen.
- Mindestens 16 Items werden benötigt, wenn möglich mehr, um nach der Erprobung eine Auswahl treffen zu können. Das erste Item dient als Beispiel und wird mit „0“ (null) nummeriert. Items sollten Präpositionen, Pronomen, Partikel, Hilfsverben etc. abprüfen. Sie sollen gleichmäßig über den Text verteilt sein, und es muss darauf geachtet werden, dass eine falsche Antwort nicht automatisch zu einer ebenfalls falschen Antwort im nächsten Item führt (Interdependenz der Items).

Die für diese Aufgabe festgelegte **Anweisung** wird dem Testautor ebenfalls vorgegeben.

Erfahrene Verfasser textbasierter Aufgaben sammeln oft kontinuierlich passende Texte aus den empfohlenen Quellen. Müssen sie dann Items erstellen, suchen sie die vielversprechendsten Texte aus dem bereits vorliegenden Material aus und arbeiten mit ihnen. Für einige Aufgabentypen (z. B. Aufgaben, die auf Grammatik oder Vokabular fokussieren) ist es sinnvoll, ein Wörterbuch oder einen Thesaurus zur Hand zu haben. Wenn Material zum Hörverstehen geschrieben wird, sollte der Text abgehört werden, so dass die Items direkt aus dem gesprochenen Text entwickelt werden und nicht aus der niedergeschriebenen Form.

Viele Testautorinnen und -autoren probieren ihre Entwürfe gerne an Kollegen oder Bekannten aus, die nichts mit Sprachtests zu tun haben. Dies hilft, um Tippfehler, unklare Anweisungen, falsche Lösungsschlüssel, sehr schwierige Items oder solche mit mehr als einer richtigen Lösung zu identifizieren.

Die **Testspezifikationen** sollten eine Checkliste enthalten, anhand derer der Testautor seinen Text, die Items und die Gesamtaufgabe vor der Abgabe überprüfen kann. Eine solche Checkliste zum oben beschriebenen C-Test wird nachfolgend als Beispiel aufgeführt. Wenn der Text, die Items und die Aufgabe angemessen sind, muss jede Frage mit „Ja“ beantwortet werden können.

Text:
Ist das Thema gebräuchlich, kulturell akzeptabel etc.?
Ist der Text frei von potentiell störendem Inhalt?
Hat er den passenden Schwierigkeitsgrad?
Eignet sich der Text für eine auf Struktur ausgerichtete Aufgabe?
Ist er lang genug, um mindestens 16 Items zu generieren?
Gibt es eine passende Überschrift?

Items:
Stimmt die erstellte Zahl der Items?
Verteilen sich die Items gleichmäßig über den Text?
Ist die sprachliche Bandbreite angemessen?
Wurde überprüft, dass sich alle Items auf Struktur beziehen?
Gibt es keine voneinander abhängigen Items?
Gibt es ein oder zwei zusätzliche Items?
Wurden eigenwillige Items vermieden?
Anweisung und Lösungsschlüssel:
Wurde die Anweisung überprüft?
Gibt es ein Beispiel?
Gibt es einen vollständigen Lösungsschlüssel auf einem separaten Blatt?

Vor dem Einreichen sollten Testautoren sicherstellen, dass sie eine Kopie ihrer Arbeit behalten haben. Besonders wenn Originaltexte aus Zeitungen oder Zeitschriften an den Testanbieter übergeben wurden, sollte der Testautor Fotokopien und Informationen zur Quelle behalten.

Anhang V: Fallstudie

Redaktionelle Bearbeitung einer Aufgabe des Niveaus B1+ einer berufsorientierten Deutsch-Prüfung

In diesem Anhang wird gezeigt, welche Veränderungen an einer Aufgabe im Zuge der redaktionellen Bearbeitung vorgenommen wurden und aus welchen Gründen dies jeweils geschah. Jede neue Version der Aufgabe wird mit entsprechenden Kommentaren zu den Änderungen präsentiert.

Version 1

Lesen Sie den Text und schließen Sie die Lücken 31–40. Benutzen Sie die Wörter a–o.
Jedes Wort passt nur einmal.
Markieren Sie Ihre Lösungen für die Aufgaben 31–40 auf dem Antwortbogen.

Judith Miller
Rüdesheimer Straße 23
65 195 Wiesbaden

Couture GmbH
Frau Erika Einsteller
Industriestr. 23
63 477 Maintal

Wiesbaden, den …

Ihre Stellenausschreibung im Internet

– Leiterin eines Friseursalons im Rhein-Main-Gebiet –

Sehr geehrte Frau Einsteller,

ich bewerbe mich um die **31** „Leiterin im Friseursalon“ bei Ihnen. Seit über sechs Jahren bin ich im Salon SchnippSchnapp in Wiesbaden als Friseurin tätig, wo ich stellvertretend bereits die **32** ausgeübt habe. Die **33** zur Friseurmeisterin habe ich soeben erfolgreich bestanden.

Alle in einem Friseursalon anfallenden **34** sind mir bestens vertraut. Dabei arbeite ich selbstständig und mit großem Engagement. Ich kann gut organisieren und gehe **35** auf die Wünsche der Kundinnen und Kunden ein. Durch meine Begeisterungsfähigkeit kann ich auch die Mitarbeiter zu hohen **36** motivieren. Selbstverständlich **37** ich mich immer über neue Trends und Techniken bei der Haarmode und **38** auch entsprechende Fortbildungsveranstaltungen. Meine **39** würde ich gerne in Ihrem Unternehmen einbringen. Die Stelle reizt mich besonders, weil ich darin die Möglichkeit sehe, mich **40** weiterzuentwickeln. Ich würde mich freuen, wenn wir uns bald über die näheren Einzelheiten der Position unterhalten könnten.

Mit freundlichen Grüßen

Judith Miller

Anlagen: Bewerbungsmappe

a	arbeiten	**d**	Fähigkeiten	**g**	Leistungen	**j**	Positionen	**m**	Stelle
b	Ausbildung	**e**	immer	**h**	Leitung	**k**	Prüfung	**n**	Tätigkeiten
c	besuche	**f**	informiere	**i**	persönlich	**l**	selbstständig	**o**	versuche

Bei der ersten Überprüfung wurde Folgendes festgestellt und an den Autor zurückgemeldet:

- Die Gestaltung der Aufgabe muss den Testspezifikationen für das entsprechende Testformat angepasst werden; alle zur Auswahl gebotenen Lösungsalternativen werden in GROSSBUCHSTABEN dargestellt. (Diese Spezifikation ist dadurch begründet, dass durch die übliche Groß-/Kleinschreibung hier nicht gewollte Hinweise zur Passung einzelner Antwortalternativen gegeben werden würden und dies einer reliablen Messung des Wortschatzes, die in diesem Testteil angestrebt wird, entgegenstehen würde.)
- In der Folge muss die Antwortalternative a „ARBEITEN“ herausgenommen werden, da dieses nach Umstellung auf reine Großschreibung für Item 34 neben „TÄTIGKEITEN“ als weitere richtige Lösung in Frage käme. (Die Testspezifikation sieht je Item nur eine mögliche richtige Lösung vor.) „ARBEITEN“ wird durch die neue Antwortalternative „BERUFLICH“ ersetzt.

Erste Redaktionssitzung

Nach Einholung von Gutachten wurde eine Redaktionssitzung abgehalten, bei der einige problematische Aspekte diskutiert wurden. Der Autor wurde um Überarbeitung folgender Punkte gebeten:

- Der Stil entspricht im einleitenden ersten Satz nicht ganz einem authentischen Bewerbungsschreiben; es sollten hier die typischen Redemittel verwendet werden.
- Auch „wo ich stellvertretend bereits die [32=Leitung] ausgeübt habe“ erscheint aufgrund fehlender Zeitangabe wenig präzise.
- Die aufgeführten Lösungsmöglichkeiten wurden eingehend analysiert. Dabei wurde erarbeitet, dass das Adverb „IMMER“ (Antwortalternative e), das als Lösung für Item 35 gedacht war, an dieser Stelle im Text unpassend wirkt. Denn es erscheint schlecht vorstellbar, dass eine Friseurin immer auf die Wünsche ihrer Kundschaft einzugehen vermag. Um eine überzeugendere Lösung für Item 35 zu haben, wird das Wort „IMMER“ ersetzt durch „GERN“.
- Der Satz, der Item 40 enthält, für welches als richtige Lösung i „persönlich“ vorgesehen ist, erscheint in mit der hier vorliegenden einseitigen Betonung der „persönlichen Weiterentwicklung“ als Motivation für die Bewerbung wenig überzeugend.

Version 2

Folgende Änderungen wurden vom Testautor vorgenommen:

- Die in der ersten Redaktionssitzung erbetenen stilistischen Anpassungen und Präzisierungen wurden vorgenommen.
- Item 40 wurde zwar nicht verändert (auch die Lösung bleibt gleich), doch wurde der vorletzte Satz des Textes so ergänzt, dass das Item nun im Kontext eines Bewerbungsschreibens authentischer und somit eindeutig lösbar erscheint.

Lesen Sie den Text und schließen Sie die Lücken 31–40. Benutzen Sie die Wörter a–o.
Jedes Wort passt nur einmal.
Markieren Sie Ihre Lösungen für die Aufgaben 31–40 auf dem Antwortbogen.

Judith Miller
Rüdesheimer Straße 23
65 195 Wiesbaden

Couture GmbH
Frau Erika Einsteller
Industriestr. 23
63 477 Maintal

Wiesbaden, den …

Ihre Stellenausschreibung im Internet

Leiterin eines Friseursalons im Großraum Rhein/Main

Sehr geehrte Frau Einsteller,
hiermit bewerbe ich mich um die **31** als Leiterin eines Friseursalons in Ihrem Unternehmen.
Seit über sechs Jahren bin ich im Salon SchnippSchnapp in Wiesbaden als Friseurin tätig, wo ich seit 2 Jahren auch die stellvertretende **32** habe. Die **33** zur Friseurmeisterin habe ich soeben erfolgreich bestanden.

Alle in einem Friseursalon anfallenden **34** sind mir bestens vertraut. Dabei arbeite ich selbstständig und mit großem Engagement. Ich kann gut organisieren und gehe **35** auf die Wünsche der Kundinnen und Kunden ein. Durch meine Begeisterungsfähigkeit kann ich auch die Mitarbeiter zu hohen **36** motivieren. Selbstverständlich **37** ich mich immer über neue Trends und Techniken bei der Haarmode und **38** auch entsprechende Fortbildungsveranstaltungen. Meine **39** würde ich gerne in Ihrem Unternehmen einbringen. Die Stelle reizt mich besonders, da ich mich beruflich, aber auch **40** weiterentwickeln möchte. Ich würde mich freuen, wenn wir uns bald über die näheren Einzelheiten der Position unterhalten könnten.

Mit freundlichen Grüßen

Judith Miller

Anlage
Bewerbungsmappe

a	AUSBILDUNG	**d**	FÄHIGKEITEN	**g**	LEISTUNGEN	**j**	POSITIONEN	**m**	STELLE
b	BERUFLICH	**e**	GERN	**h**	LEITUNG	**k**	PRÜFUNG	**n**	TÄTIGKEITEN
c	BESUCHE	**f**	INFORMIERE	**i**	PERSÖNLICH	**l**	SELBSTSTÄNDIG	**o**	VERSUCHE

Besprechung der erprobten Aufgabe

Die Aufgabe wurde nun erprobt. In der nächsten Redaktionssitzung wurden die Daten aus einer statistischen Item-Analyse nach der Klassischen Testtheorie herangezogen (zur Interpretation der in der nachfolgenden Tabelle berichteten Statistiken siehe Anhang VII). Diese ergab eine recht gute Streuung der Schwierigkeitswerte. Item 32 fällt jedoch mit einem sehr niedrigen Wert (p=.14) auf, der eine hohe Schwierigkeit indiziert. Acht der zehn Items weisen hervorragende Trennschärfeindizes auf. Ausnahmen bilden hier lediglich die Items 33 (rpbis =.21) und 40 (rpbis =.28).

Item Statistics					Alternative Statistics					
Seq. No.	Scale-Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
31	2-11	.83	.46	.54	1	.83	.54	1.00	.54	*
					2	.00	.00	.00		
					Other	.17	.00	.00	-.54	
32	2-12	**.14**	.38	57	1	.14	.00	.38	.57	*
					2	.00	.00	.00		
					Other	.86	.00	.00	-.57	
33	2-13	.74	.23	**.21**	1	.74	.69	.92	.21	*
					2	.00	.00	.00		
					Other	.26	.00	.00	-.21	
34	2-14	.31	.38	.44	1	.31	.23	.62	.44	*
					2	.00	.00	.00		
					Other	.69	.00	.00	-.44	
35	2-15	.76	.46	.55	1	.76	.38	.85	.55	*
					2	.00	.00	.00		
					Other	.24	.00	.00	-.55	
36	2-16	.45	.92	.74	1	.45	.00	.92	.74	*
					2	.00	.00	.00		
					Other	.55	.00	.00	-.74	
37	2-17	.60	.62	.58	1	.60	.31	.92	.58	*
					2	.00	.00	.00		
					Other	.40	.00	.00	-.58	
38	2-18	.55	.85	.65	1	.55	.00	.85	.65	*
					2	.00	.00	.00		
					Other	.45	.00	.00	-.65	
39	2-19	.45	.85	.67	1	.45	.00	.85	.67	*
					2	.00	.00	.00		
					Other	.55	.00	.00	-.67	
40	2-20	.33	.31	**.28**	1	.33	.15	.46	.28	*
					2	.00	.00	.00		
					Other	.67	.00	.00	-.28	

Abb. 18: Erprobungsergebnisse (statistische Itemanalyse nach der Klassischen Testtheorie)

Im Zuge der erneuten inhaltlichen Analyse und Besprechung dieser drei Items mit auffälligen Kennwerten wurden folgende Erklärungsmöglichkeiten für die ungünstige Performanz diskutiert und sich daraus ergebende Überarbeitungsansätze beschlossen:

- Item 32: Hier könnte die sprachlich nicht gut gelungene Konstruktion „Leitung haben“ dazu beigetragen haben, dass das Item von vielen Teilnehmern nicht gelöst werden konnte. Es wurde folgende Umformulierung beschlossen: „…, wo ich seit 2 Jahren auch die stellvertretende **32** des Salons habe.“
- Item 33: Als mögliche Interpretation des unbefriedigenden Trennschärfewertes wurde diskutiert, dass vielen Erprobungsteilnehmern, die dieses Item eigentlich hätten lösen können müssen, unter Umständen ein Flüchtigkeitsfehler unterlaufen sein könnte. Wenn sie den Satz vor der Lösung des Items nicht sorgfältig bis zum Ende gelesen haben, wurde eventuell öfter „Prüfung bestehen“ mit „Ausbildung machen“ verwechselt. Eine weitere mögliche Schwierigkeit liegt mutmaßlich darin, dass die berufliche Weiterqualifizierung zur Friseurmeisterin bei einigen Teilnehmern mit „Ausbildung“ in Verbindung gebracht worden sein könnte. Es wurde daher entschieden, das Item wie folgt zu ändern: „Die Prüfung zur Friseurmeisterin habe ich soeben **33**.“ Die jetzigen Antwortmöglichkeiten „AUSBILDUNG“ (a) und „PRÜFUNG“ (k) werden ersetzt durch „BESTANDEN“ und „ANGENOMMEN“ (letzteres als Distraktor für die richtigen Lösung „BESTANDEN“).
- Item 40: Es wurde diskutiert, dass die schlechte Trennschärfe dieses Items möglicherweise darin begründet sein könnte, dass das Konzept der „persönlichen Weiterentwicklung“ zumindest für Menschen aus anderen Kulturkreisen als unpassend erachtet werden könnte; entsprechend würden diese die Antwortalternative i „PERSÖNLICH“ nicht als Lösung für Item 40 in Betracht ziehen. Es wurde daher folgende Änderung beschlossen: Die Lücke für Item 40 wird weiter vorne im Satz gesetzt, und zwar an die Stelle des Wortes „BERUFLICH“. Dieses Wort, das somit die neue richtige Lösung des Items 40 darstellt, ist bereits in den dargebotenen Antwortalternativen enthalten (Alternative b). Eine Änderung des Lösungsschlüssels wurde damit erforderlich. Zudem wurde eine neue Antwortalternative i benötigt. Denn durch die Verschiebung der Lücke für Item 40 steht das bisherige „PERSÖNLICH“ nun im Text. Als neue Alternative i wurde das Wort „PRIVAT“, das in diesem Zusammenhang als guter Distraktor für die richtige Antwort b „BERUFLICH“ erscheint, gewählt.

Version 3

Die entsprechend überarbeitete Version der Aufgabe ist noch nicht „reif“ für den Einsatz in einer echten Prüfung. Es besteht weiterer Erprobungsbedarf; insbesondere ist zu überprüfen, ob durch die vorgenommenen Modifikationen tatsächlich die gewünschten Verbesserungen der Itemschwierigkeit (Item 32) bzw. Trennschärfe (Items 33 und 40) erzielt werden konnten.

Lesen Sie den Text und schließen Sie die Lücken 31–40. Benutzen Sie die Wörter a–o.
Jedes Wort passt nur einmal.
Markieren Sie Ihre Lösungen für die Aufgaben 31–40 auf dem Antwortbogen.

Judith Miller
Rüdesheimer Straße 23
65 195 Wiesbaden

Couture GmbH
Frau Erika Einsteller
Industriestr. 23
63 477 Maintal

Wiesbaden, den …

Ihre Stellenausschreibung im Internet

Leiterin eines Friseursalons im Großraum Rhein/Main

Sehr geehrte Frau Einsteller,

hiermit bewerbe ich mich um die **31** als Leiterin eines Friseursalons in Ihrem Unternehmen. Seit über sechs Jahren bin ich im Salon SchnippSchnapp in Wiesbaden als Friseurin tätig, wo ich seit 2 Jahren auch die stellvertretende **32** des Salons habe. Die Prüfung zur Friseurmeisterin habe ich soeben **33**.

Alle in einem Friseursalon anfallenden **34** sind mir bestens vertraut. Dabei arbeite ich selbstständig und mit großem Engagement. Ich kann gut organisieren und gehe **35** auf die Wünsche der Kundinnen und Kunden ein. Durch meine Begeisterungsfähigkeit kann ich auch die Mitarbeiter zu hohen **36** motivieren. Selbstverständlich **37** ich mich immer über neue Trends und Techniken bei der Haarmode und **38** auch entsprechende Fortbildungsveranstaltungen. Meine **39** würde ich gerne in Ihrem Unternehmen einbringen. Die Stelle reizt mich besonders, da ich mich **40**, aber auch persönlich weiterentwickeln möchte. Ich würde mich freuen, wenn wir uns bald über die näheren Einzelheiten der Position unterhalten könnten.

Mit freundlichen Grüßen

Judith Miller

Anlage
Bewerbungsmappe

a	ANGENOMMEN	**d**	BESUCHE	**g**	INFORMIERE	**j**	POSITIONEN	**m**	STELLE
b	BERUFLICH	**e**	FÄHIGKEITEN	**h**	LEISTUNGEN	**k**	PRIVAT	**n**	TÄTIGKEITEN
c	BESTANDEN	**f**	GERN	**i**	LEITUNG	**l**	SELBSTSTÄNDIG	**o**	VERSUCHE

Anhang VI: Informationen aus Erprobungen

Dieser Anhang zeigt mögliche Fragen auf, die nach dem **Vorerproben** und der **Erprobung** (siehe Kapitel 3.4.2) gestellt werden sollten.

Rückmeldungen der Aufsichtspersonen – alle Teile

Bitte kommentieren Sie folgende Aspekte:

- Inhalt: Bandbreite und Art der Texte und Aufgaben
- Niveau: Schwierigkeit (z. B. sprachliche oder kognitive) der verschiedenen Testteile und Aufgaben
- Nur Hörverstehen: Klarheit und Geschwindigkeit des Gesprochenen, Akzent der Sprecher etc.
- Teilnehmende: Was ist das ungefähre Alter der Erprobungsteilnehmer?
- Weitere Kommentare?

Rückmeldungen der Erprobungsteilnehmer – Leseverstehen

- Hatten Sie ausreichend Zeit zur Bearbeitung der Aufgaben? (Wenn nicht, wie viel mehr Zeit hätten Sie benötigt?)
- Gab es in der Aufgabe Vokabular, das Sie nicht verstanden haben? (Bitte vermerken Sie Wörter oder Ausdrücke, die problematisch waren.)
- Konnten Sie den Ideen und der Argumentationsstruktur des Textes folgen? (leicht/mit einigen Schwierigkeiten/mit großen Schwierigkeiten)
- Wie vertraut waren Sie mit dem Thema des Texts? (sehr vertraut/ziemlich vertraut/nicht sehr vertraut/gar nicht vertraut)
- Wann (wenn überhaupt) planen Sie, die Prüfung tatsächlich abzulegen?
- Möchten Sie weitere Kommentare machen?

Rückmeldungen aus der Bewertung – Schreibaufgabe

Zur Aufgabe: Aufgabenstellung

- Wurde die Aufgabe allgemein verstanden?
- Wurde die Rolle des Schreibers klar verstanden?
- Wurde der Adressat klar identifiziert?
- Gibt es eine Verzerrung der Ergebnisse (*Bias*)? Haben Teilnehmende mit bestimmtem kulturellen Hintergrund oder aus einer bestimmten Altersgruppe Vorteile?
- Muss der Wortlaut der Aufgabe geändert werden? Wenn ja, machen Sie bitte Vorschläge.

Zur Aufgabe: Sprache

- Ist die Aufgabe für Teilnehmende auf der angezielten GER-Stufe verständlich?
- Verursachte die Formulierung der Aufgabe Irritationen oder Missverständnisse?
- War den Teilnehmenden das passende **Register** klar?
- Muss der Wortlaut der Frage geändert werden? Wenn ja, machen Sie bitte Vorschläge.

Zur Schreibleistung: Inhalt

- Wurde die Aufgabe in angemessener Weise verstanden?
- Wurde ein Punkt zum Inhalt falsch interpretiert oder ausgelassen? Bitte kommentieren.
- War die Textlänge für die Aufgabe angemessen?

Zur Schreibleistung: Spektrum/Register

- Wurden sprachliche Mittel aus der Aufgabe übernommen? Bitte kommentieren.
- Welches Register (formal, persönlich etc.) haben die Teilnehmenden verwendet?

Zur Schreibleistung: Niveau

- Gab die Aufgabe den Teilnehmenden genügend Spielraum, um ihre Fähigkeiten zu zeigen?

Zu den Bewertungskriterien:

- Bitte vermerken Sie Vorschläge zu Änderungen der Bewertungskriterien.

Allgemeiner Eindruck:

- Bitte vermerken Sie Ihren allgemeinen Eindruck hinsichtlich der vorliegenden Aufgabe.

Anhang VII: Statistische Analysen

Daten zu einem Test zu erheben und zu analysieren, erfordert Planung und Ressourcen, trägt aber entscheidend zur Verbesserung der Testqualität und der Interpretierbarkeit der Ergebnisse bei. Als Minimallösung müssen Angaben zu den Teilnehmenden und zu ihren Ergebnissen bzw. Noten festgehalten werden. Diese Angaben können dann mit einfachen statistischen Methoden analysiert werden (vgl. z. B. Carr 2008).

Genauere Erhebungen zu Teilnehmerleistungen können zeigen, wie gut Items funktioniert haben und worauf man sich bei der redaktionellen Bearbeitung konzentrieren sollte. Es gibt benutzerfreundliche Software-Pakete für die unten beschriebenen Analysezwecke – auch für eine kleine Anzahl von Prüfungsteilnehmerinnen und -teilnehmern (z. B. 50).

Zu erhebende Daten sind u. a:

- Daten auf Aufgabenebene: Wie war das Ergebnis der Teilnehmenden in jeder einzelnen Aufgabe und nicht nur insgesamt?
- Daten auf Itemebene: Wie haben Teilnehmende jedes einzelne Item gelöst?
- Demografische Informationen über die Teilnehmenden: Alter, Geschlecht, Erstsprache etc.

Die Daten

Der Großteil der Software für die Klassische Analyse erfordert Daten wie in Abbildung 19 gezeigt. Man kann jedes Textverarbeitungsprogramm zur Eingabe der Daten verwenden, muss aber auf Folgendes achten:

- eine Schriftart mit einer festgelegten Breite, also eine Nicht-Proportionalschrift wie Courier wählen
- keine Tabulatoren benutzen
- die Datei als Textdokument (.txt) speichern

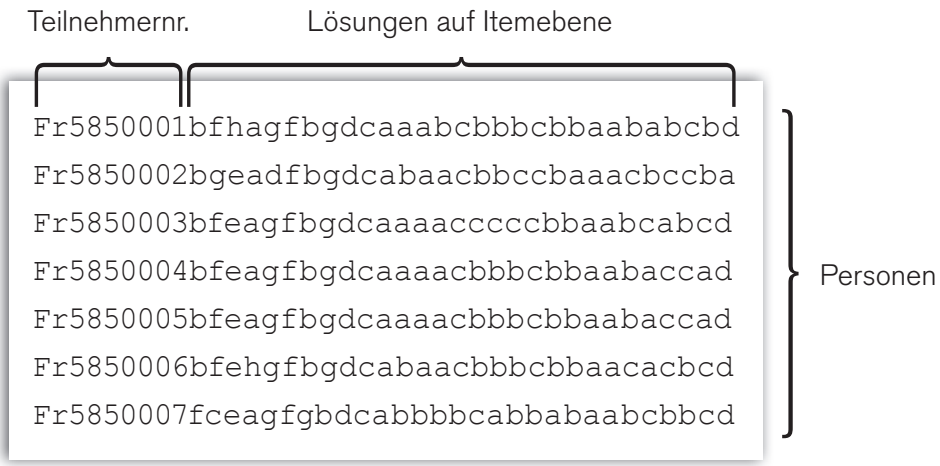


Abb. 19: Eine typische Datei zur Datenerhebung auf Itemebene

In Abbildung 19 sind die Daten angeordnet wie folgt:

- Jede Zeile zeigt die Lösungen einer Person.
- Die ersten Spalten geben die Identifikationsnummer für die jeweilige Person an (diese kann auch demografische Informationen enthalten).
- Jede weitere Spalte zeigt die Lösung eines Items.

Dieses Beispiel zeigt einen Multiple-Choice-Test, in dem die jeweilige Wahl (a–h) eines jeden Teilnehmenden festgehalten wird.

Die Analyse-Software benötigt zusätzliche Informationen, u. a. die richtige Lösung für das jeweilige Item.

Klassische Itemanalyse

Die Klassische **Itemanalyse** wird eingesetzt wie folgt:

- zur Analyse von Daten aus Erprobungen, um Angaben über die Auswahl und die Redaktion der Aufgaben für den Echteinsatz zu machen
- zur Analyse der Daten aus dem Echteinsatz

Sie erhebt verschiedene statistische Angaben auf Ebene der Items und des Gesamttests, vor allem die folgenden:

Daten zum Verhalten jedes einzelnen Items:

- wie leicht das Item für eine Teilnehmergruppe war
- wie scharf das Item zwischen stärkeren und schwächeren Teilnehmenden trennt
- wie gut die Lösung und jeder Distraktor funktioniert

Zusammenfassende Statistiken zum Gesamttest oder zu Testeilen:

- Anzahl der Teilnehmerinnen und Teilnehmer
- Mittelwert und **Standardabweichung** der Ergebnisse
- Einschätzung zur Reliabilität

Nachfolgend geben wir einige Angaben dazu, welche Größen für diese Statistiken akzeptabel sind. Sie stellen keine Regeln dar – in der Praxis hängen die normalerweise beobachteten Größen vom Kontext ab. Methoden der Klassischen Statistik ergeben unter folgenden Voraussetzungen die besten Ergebnisse:

- große Anzahl von Items in einem Test
- viele Prüflingsteilnehmerinnen und -teilnehmer
- eine größere Bandbreite von Fähigkeiten innerhalb der Teilnehmergruppe

Umgekehrt sind die Ergebnisse der Analyse weniger aussagekräftig, wenn es nur wenige Items oder Teilnehmende gibt und alle ungefähr die gleichen Fähigkeiten haben.

Abbildung 19 gibt ein Beispiel zur Itemstatistik auf der Grundlage der Ergebnisse der MicroCAT-Analyse-Software (s. u. bei Hinweisen zu Analysesoftware). Sie zeigt die Analyse für drei Items.

Item Statistics					Alternative Statistics					
Seq. No.	Scale-Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
1	1–1	.38	.52	.48	A	.00	.00	.00		
					B	.38	.13	.66	.48	*
					C	.12	.11	.12	–.01	
					D	.49	.74	.23	.44	
					Other	.01	.00	.00	–.11	
2	1–1	.38	.52	.48	A	.00	.00	.00		
					B	.38	.13	.66	.48	*
					C	.12	.11	.12	–.01	
					D	.49	.74	.23	.44	
					Other	.01	.00	.00	–.11	
3	1–1	.38	.52	.48	A	.00	.00	.00		
					B	.38	.13	.66	.48	*
					Other	.01	.00	.00	–.11	

Abb. 20: Beispiel einer Itemstatistik (MicroCAT item analysis package)

Schwierigkeit

Der **Schwierigkeitswert** bezeichnet den Anteil der richtigen Lösungen (*Prop. correct* in Abbildung 20). Er zeigt, wie einfach ein Item für die jeweilige Gruppe von Prüfungsteilnehmerinnen und -teilnehmern war. Der Wert liegt zwischen 0 und 1; je höher die Zahl, desto leichter das Item. Abbildung 20 zeigt, dass Item 1 das schwierigste, Item 3 das einfachste war.

Der Schwierigkeitswert sollte in einer Statistik immer zuerst betrachtet werden, denn wenn er zu hoch oder zu niedrig ist (z. B. außerhalb der Spannweite 0.25–0.80), so bedeutet dies, dass andere Statistiken keine angemessenen Rückschlüsse ermöglichen. Von der geprüften Gruppe kommen keine sinnvollen Informationen; oder, wenn die untersuchte Gruppe die Teilnehmenden der Echtprüfung repräsentiert, dann kann man schlussfolgern, dass das Item einfach zu leicht oder zu schwierig war. Ist das Niveau der Teilnehmenden nicht sicher, so ist vielleicht das Item gut, aber die Teilnehmenden hatten nicht das richtige Sprachniveau. Grundsätzlich kann man festhalten, dass ein Test immer mit einer Gruppe erprobt werden sollte, die das gleiche Niveau hat wie die tatsächlichen Prüfungsteilnehmenden.

Trennschärfe

Gute Items unterscheiden angemessen zwischen schwächeren und stärkeren Prüfungsteilnehmenden. Klassische Itemanalysen bieten zwei Indizes hierfür: den **Trennschärfe**-Index und die punktbiseriale **Korrelation** (*Disc. Index* und *Point Biser.* in Abbildung 20).

Der Trennschärfeindex ist eine einfache Statistik: Sie ergibt sich aus der Differenz zwischen dem Anteil richtiger Lösungen bei Teilnehmenden mit den besten Ergebnissen und denjenigen mit den schwächsten Ergebnissen (normalerweise das oberste und unterste Drittel der Gruppe). Das Ergebnis wird in Abbildung 20 in den Spalten *low* (niedrig) und *high* (hoch) dargestellt. Bei Item 1 liegt der Unterschied zwischen der stärkeren und der schwächeren Gruppe bei (0.66–0.13). Dies ist der Wert des Trennschärfeindexes (Rundungsfehler sind möglich).

Ein hoher Trennschärfewert für ein Item wäre +1. Er würde zeigen, dass die besten Teilnehmenden dieses Item immer richtig lösen, während die schwächeren eine falsche Lösung wählen.

Ist der Schwierigkeitswert entweder sehr hoch oder sehr niedrig, heißt das, dass sowohl die stärksten als auch die schwächsten Teilnehmenden das Item richtig lösen (oder falsch) und so die Trennschärfe für dieses Item also falsch eingeschätzt wird. Item 3 zeigt dieses Problem: 1.00–0.81 = 0.19 ist ein niedriger Wert.

Der punktbiseriale Index bietet eine komplexere Berechnung als der Trennschärfenindex und hält einem hohen oder niedrigen Schwierigkeitswert besser stand. Er stellt eine Korrelation zwischen den Ergebnissen der Teilnehmer für ein Item (1 oder 0) und für den gesamten Test dar.

Allgemein gilt, dass Items mit einer punktbiserialen Korrelation von mehr als 0.30 akzeptabel sind. Ein negativer Wert heißt, dass gute Teilnehmende dieses Item eher falsch beantworten. In diesem Fall sollte überprüft werden, ob einer der Distraktoren tatsächlich die richtige Lösung ist oder ob der Lösungsschlüssel falsch ist.

Distraktorenanalyse

Distraktoren sind die falschen Optionen bei einem Multiple-Choice-Item. Es ist zu erwarten, dass schwächere Teilnehmende eher den Distraktor wählen, stärkere dagegen die korrekte Lösung laut Schlüssel („Key“, markiert mit „*“).

Eine Distraktorenanalyse zeigt, welcher Anteil der Teilnehmenden welchen Distraktor gewählt hat (*Prop. Total* in Abbildung 20). Item 1 in Abbildung 20 hat einen ziemlich niedrigen Schwierigkeitswert für die korrekte Option B: 0.38. Distraktor D erhält mehr Antworten (Schwierigkeitswert = 0.49). Distraktor A dagegen wird überhaupt nicht gewählt, er ist also kein guter Distraktor. Allgemein funktioniert das Item jedoch gut, mit einer angemessenen Trennschärfe, so dass es nicht geändert werden muss. Die Praxis zeigt, dass es schwierig ist, drei Distraktoren zu finden, die gleich gut funktionieren.

Die Analyse in Abbildung 20 zeigt außerdem die Wahl des stärkeren und des schwächeren Teilnehmeranteils sowie den punktbiserialen Wert für jede Option. Ein gutes Item wird einen positiven Wert für die korrekte Lösung und einen negativen Wert für jeden Distraktor erzielen.

Reliabilität der Ergebnisse

Reliabilität kann auf verschiedene Weise geschätzt und mit verschiedenen Formeln berechnet werden. Jede Methode geht von unterschiedlichen Annahmen aus. Bei der Halbierungsmethode wird der Test in zwei gleich große Teile geteilt, und das Ergebnis desselben Teilnehmers in beiden Teilen wird verglichen. Bei dieser Methode ist es wichtig, dass die beiden Teile so ähnlich wie möglich sind, vor allem durch Übereinstimmung im Konstrukt und beim Schwierigkeitsgrad.

Andere Methoden messen die interne Konsistenz eines Tests. Dies funktioniert dann gut, wenn die Items sehr ähnlich in Typ und Inhalt sind. Handelt es sich jedoch um recht unterschiedliche Items, wird die Reliabilität als zu niedrig eingeschätzt.

Zur klassischen Itemanalyse:

Mindestanzahl der Teilnehmenden: 50 bis 80 (Jones, Smith und Talley 2006: 495)
Weitere Informationen: Verhelst (2004a, b); Bachman (2004)

Rasch-Analyse

Die **Rasch-Analyse** ist die einfachste und praktikabelste Form des **probabilistischen Modells** (oder Item-Response-Theorie, IRT). Sie erleichtert das Verständnis für Itemschwierigkeit im Vergleich zur Klassischen Itemanalyse und bietet zusätzliche Anwendungen, wie z. B. die **Verknüpfung** von Testversionen.

Mit der Rasch-Analyse

- wird der genaue Unterschied im Schwierigkeitsgrad zwischen zwei Items klar, da die Items auf einer **Intervallskala**, gemessen in Logits, platziert werden,
- kann der Unterschied zwischen Items und Teilnehmenden, Testergebnissen und Bestehensgrenzen einheitlich verdeutlicht werden, da alle diese Punkte auf einer einzigen Skala abgetragen werden,
- kann die Schwierigkeit der Items unabhängig von der Fähigkeit der Teilnehmenden gezeigt werden (die Klassische Itemanalyse kann dazu führen, dass eine starke Probandengruppe ein Item leicht erscheinen lässt, wogegen eine schwächere Gruppe dasselbe Item schwierig erscheinen lässt).

Aufgrund dieser Möglichkeiten ist die Rasch-Analyse gut geeignet, um die Standards bei jedem Prüfungsereignis zu überwachen und aufrechtzuerhalten. Um eine solche Rasch-Analyse durchführen zu können, muss man die verschiedenen Testversionen zunächst miteinander verknüpfen. Dazu stehen mehrere Methoden zur Verfügung:

- einige identische Items in beiden Testversionen verwenden
- bei beiden Prüfungsereignissen zusätzlich **Ankeritems** einsetzen
- einige oder alle Items kalibrieren, bevor sie in den Echteinsatz kommen (siehe Kapitel 3.4.2, Erprobung)
- einigen Teilnehmenden beide Testversionen vorlegen

Bei der Analyse beider Testversionen bietet deren Verknüpfung nun einen übergreifenden Rahmen für die Einordnung aller Items, Prüfungsteilnehmerinnen und -teilnehmer usw., und die Items erhalten kalibrierte Schwierigkeitswerte. Weitere Testversionen können diesen Rahmen nun auf die gleiche Art und Weise erweitern.

Standards überwacht man, indem die jeweilige Position der wichtigen Elemente verglichen werden:

- Sind die Items in allen Testversionen auf dem gleichen Schwierigkeitsniveau?
- Haben die Teilnehmenden die gleichen Fähigkeiten?
- Stimmen die Bestehensgrenzen (gemessen in Logits) in allen Testversionen mit den gleichen **Rohwerten** überein (jetzt auch in Logits gemessen)?

Wenn die Bestehensgrenzen jedes Mal auf dem gleichen Schwierigkeitsniveau liegen, ist die Standardisierung erfolgreich.

Es ist leichter, Qualität und Standards zu sichern, wenn Tests mit **kalibrierten** Items erstellt werden. Die Schwierigkeit der gesamten Testversion kann anhand seiner Durchschnittsschwierigkeit und der Spannweite von Schwierigkeitswerten beschrieben werden. Sie kann gelenkt werden, indem man Items auswählt,

die zusammengenommen der angezielten Spannweite der Schwierigkeit und dem gewünschten Mittelwert entsprechen.

Wenn man beginnt, Items zu kalibrieren, werden die Schwierigkeitswerte zunächst nicht viel aussagen Im Laufe der Zeit aber kann man beobachten, was Prüfungsteilnehmende tatsächlich können und so den Punkten auf der Schwierigkeitsskala eine Bedeutung zuweisen. Alternativ kann man anschließend oder gleichzeitig auch subjektive Beurteilungen über die Items anstellen („Ich denke, ein Lerner an der B1-Grenze hat eine 60-prozentige Chance, dieses Item richtig zu beantworten."), um die Schwierigkeitswerte zu interpretieren. So werden die Werte auf der Skala vertraut und aussagekräftig.

Zur Rasch-Analyse:

Mindestanzahl der Teilnehmenden: 50 bis 80 (Jones, Smith und Talley 2006: 495)
Weitere Informationen: Verhelst (2004d); Bond und Fox (2007)

Statistiken zur Auswertung und Bewertung

Manuelle Auswertung

Es ist wichtig festzustellen, ob bei der Auswertung gute Arbeit geleistet wird. Ist die Leistung des Auswertungspersonals nicht zufriedenstellend, können Maßnahmen – wie z. B. eine erneute Schulung – ergriffen werden (siehe Kapitel 5.1). Bei kleinen Prüfungsgruppen kann man die Entscheidung eines jeden Auswerters zu jedem Item überprüfen. Bei einer größeren Gruppe sollte eine Stichprobe (z. B. 10 %) der Auswertungen kontrolliert und eine Fehlerquote errechnet werden. Eine Fehlerquote bezeichnet die Anzahl der Auswertungsfehler geteilt durch die Anzahl der auszuwertenden Items. Spiegelt die Stichprobe die gesamte Arbeit des jeweiligen Auswerters angemessen wider, so wird die Fehlerquote immer die gleiche sein.

Damit die Auswertungsstichprobe repräsentativ für die Arbeit der jeweiligen Person ist, sollte sie am besten zufällig ausgewählt werden. Dafür ist es wichtig, die Arbeitsweise nachvollziehen zu können. Eine Zufallsstichprobe ist nicht gleichzusetzen mit der Auswahl irgendwelcher 10 Prozent der Testunterlagen, denn so sind möglicherweise nur die letzten Arbeiten in der Auswahl enthalten, da diese leichter zugänglich sind. Somit wäre die Fehlerquote vielleicht für die gesamte Arbeitszeit zu niedrig angesetzt; die Auswertungsleistung verbesserte sich sicherlich mit der Zeit.

Bewertung

Die Leistung der Bewerterinnen und **Bewerter** kann statistisch ganz einfach überprüft werden, indem der Durchschnittswert ihrer Bewertungen und die **Standardabweichung** (ein Maß für die Streuung ihrer Bewertungen von niedrig bis hoch) berechnet wird. Man vergleicht die Bewerter miteinander, so dass diejenigen, die sich stark von den anderen unterscheiden, dann noch einmal überprüft werden können. Dies funktioniert, wenn die Prüfungsmaterialien nach dem Zufallsprinzip zur Bewertung verteilt werden. Ist dies nicht der Fall, so erhält eine Person möglicherweise Bewertungsaufträge für erwartungsgemäß stärkere oder schwächere Gruppen. Der Durchschnittswert wird dann höher oder niedriger als bei den anderen Bewertern sein, aber die Leistung dieses ausgewählten Bewerters ist dennoch gut.

Werden Aufgaben von zwei Bewerterinnen oder Bewertern bewertet, kann die Reliabilität dieser Bewertungen beispielsweise in Excel mit dem „Pearson"-Korrelationskoeffizienten überprüft werden. Die Daten werden folgendermaßen aufgebaut:

	Bewerter 1	Bewerter 2
Teilnehmer 1	5	4
Teilnehmer 2	3	4
Teilnehmer 3	4	5
…	…	…

Der Korrelationskoeffizient liegt zwischen – 1 und 1. Im Allgemeinen sollte man jedem Wert unter 0,8 genauer nachgehen, da dies auf eine unterschiedliche Bewertungsleistung hindeutet.

Ein Reliabilitätswert, wie z. B. der ALPHA-Wert einer MicroCAT-Analyse (siehe nachfolgende Werkzeuge für die statistische Analyse), kann für alle Bewerter berechnet werden. Die Daten können etwa wie in Abbildung 20 präsentiert werden: Jede Reihe zeigt die Leistung eines Prüfungsteilnehmenden bei einer Aufgabe; jede Spalte zeigt die Bewertungen.

Die Multifacetten-Rasch-Analyse *(Many-Facet Rasch Measurement, MFRM)*

Eine anspruchsvollere Art Bewertungsleistungen zu überprüfen, stellt etwa die **Multifacetten-Rasch-Analyse** dar. Dies ist eine Variante der Rasch-Analyse und kann mit der FACETS-Software gerechnet werden (Linacre 2009). Die Analyse misst genau wie im Rasch-Modell die Schwierigkeit der Aufgaben und die Fähigkeit der Prüfungsteilnehmenden, kann zusätzlich aber die Strenge und Milde der Bewerterinnen und Bewerter überprüfen. Somit lassen sich faire Ergebnisse für die Teilnehmenden erzielen, da Strenge- und Mildeeffekte ausgeglichen werden können.

Bei der Durchführung dieser Analyse muss man darauf achten, die Daten der Bewerter und Teilnehmer, der Aufgaben und anderer gemessener Facetten miteinander zu verknüpfen. So müssen z. B. einige Teilnehmerleistungen von mehr als einer Person bewertet werden, um die Verknüpfung zwischen den Bewertern sicherzustellen. Einige Teilnehmende müssen mehr als eine Aufgabe lösen, um eine Verbindung zwischen den Aufgaben herzustellen. Wenn Daten unabhängig voneinander erhoben werden, kann die Multifacetten-Rasch-Analyse keine Überprüfung aller Elemente sicherstellen.

Zum Multifacetten-Rasch-Modell (MFRM):

Mindestanzahl der Teilnehmerleistungen: 30 für jede zu bewertende Aufgabe (Linacre 2009)
Mindestanzahl der Bewertungen von jedem Bewerter: 30 (Linacre 2009)
Weitere Informationen: Eckes (2009)

Konstruktvalidität

Überprüfen der Teststruktur

Faktorenanalysen oder Strukturgleichungsmodelle können zeigen, ob die Items eines Tests das beabsichtigte Konstrukt treffen. Der Test sollte ein Muster aufzeigen, das mit dem zugrundeliegenden Modell zur Sprachverwendung (siehe Kapitel 1.1) übereinstimmt. Faktorenanalysen sind für die Phase der Testentwicklung nützlich zum Abgleich, ob der Test bzw. dessen Spezifikationen wie erwartet funktionieren.

Zur Faktorenanalyse:

Mindestanzahl von Teilnehmenden: 200 (Jones, Smith und Talley 2006: 495)
Weitere Informationen: Verhelst (2004C)

Aufdecken von Verzerrung

Verzerrung (*bias*) tritt dann auf, wenn Items auf unfaire Weise bestimmte Gruppen von Prüfungsteilnehmerinnen und -teilnehmern mit den gleichen Fähigkeiten bevorzugen oder benachteiligen. So könnte beispielsweise ein Item für Teilnehmerinnen leichter sein als für Teilnehmer, obwohl beide die gleichen sprachlichen Fähigkeiten haben. In dem Fall ist der Test nicht fair, denn er soll ja die Unterschiede in der Sprachfähigkeit testen und nicht zwischen den Geschlechtern differenzieren (siehe Kapitel 1.4).

Wenn eine solche Verzerrung aufgedeckt wird, gibt es bestimmte Dinge zu beachten, denn nicht alle Unterschiede zwischen Gruppen sind unfair. So können Lerner aufgrund unterschiedlicher Erstsprachen ein Item in der Zielsprache unterschiedlich schwierig finden. Wenn Sprachkompetenz gemessen wird, muss ein solcher Unterschied als Teil der Sprachkompetenz in der Zielsprache akzeptiert werden und darf nicht als Problem bei der Beurteilung gelten.

Ein Ansatz zur Reduzierung von Verzerrungen ist die Methode des so genannten *Differential Item Functioning* (DIF), um eine Verzerrung zunächst aufzudecken und diese im Anschluss weiter zu untersuchen. Hier werden die Teilnehmer-Lösungen innerhalb von Gruppen mit gleichen Fähigkeiten verglichen. Wenn der Test beispielsweise für Erwachsene aller Altersgruppen gedacht ist, dann vergleicht man die Leistungen von jüngeren und älteren Erwachsenen mit den gleichen Fähigkeiten (gemäß dem Test). Analysen, die auf der probabilistischen Testtheorie basieren, sind für diesen Ansatz gut geeignet.

Zur DIF-Analyse mit Rasch-Analyse:

Mindestanzahl von Teilnehmenden: 500, mit jeweils mindestens 100 pro Gruppe (Jones, Smith und Talley 2006: 495)
Weitere Informationen: Camilli und Shepard (1994); Clauser und Mazor (1998)

Überprüfung einer Teilnehmerstichprobe

Bei der Verwendung von Teilnehmerdaten zu Analyse und Forschungszwecken müssen die Daten im Normalfall repräsentativ für die geprüfte Gruppe (Population) sein. Informationen über Prüfungsteilnehmerinnen und -teilnehmer sollten regelmäßig gesammelt und überprüft werden. So kann man entscheiden, ob die Analyse auf der Grundlage einer tatsächlich repräsentativen Auswahl von Teilnehmenden erfolgt.

Daten über die Eigenschaften der Teilnehmerinnen und Teilnehmer können bei jeder Prüfungsdurchführung gesammelt werden (siehe Kapitel 4). Diese Eigenschaften können dann aufgrund einfacher Prozentrechnung verglichen werden, etwa zum Vergleich des Männer- und Frauenanteils in zwei unterschiedlichen Stichproben.

Ein weiter verfeinerte Analyse wird zudem untersuchen, ob Unterschiede in den Stichproben zufällig sind. Hierfür eignet sich ein Chi-Quadrat-Test. Die Ergebnisse einer solchen Analyse müssen dann qualitativ überprüft werden, um festzustellen, ob solche Unterschiede voraussichtlich zu erheblichen Leistungsunterschieden im Test führen.

Instrumentarium zur Statistischen Analyse

Es gibt verschiedene kommerzielle Softwarepakete für die Durchführung der oben genannten Analysen. Viele Berechnungen können ganz einfach mit Microsoft Excel oder anderen allgemein bekannten Programmen zur Tabellenkalkulation durchgeführt werden. Spezialisierte Anbieter stellten Software-Lösungen für verschiedene Analyseformen zur Verfügung. Sie werden unten in alphabetischer Reihenfolge aufgeführt. In einigen Fällen sind Demo-Fassungen oder Testversionen für Schüler und Studenten verfügbar.

Assessment Systems	http://www.asses.com/softwarebooks.php
Curtin University of Technology	http://lertap.curtin.edu.au/index.htm
RUMM Laboratory	http://www.rummlab.com.au
Winsteps	http://www.winsteps.com/index.htm

Einige kostenlose Programme für spezielle Zwecke:

William Bonk, University of Colorado	http://psych.colorado.edu/~bonk/
Del Siegle, University of Connecticut	http://www.gifted.uconn.edu/siegle/research/Instrument%20Reliability%20and%20Validity/Reliability/reliabilitycalculator2.xls

Anhang VIII: Glossar

Ankeritem Ein Item, das in zwei oder mehreren Tests enthalten ist. Ankeritems haben bekannte Eigenschaften oder Kennwerte und bilden einen Teil einer neuen Testversion. Dadurch sind Informationen über die neue Version verfügbar und über die Prüfungsteilnehmerinnen und -teilnehmer, die sie bearbeiten. Mit Hilfe von Ankeritems, deren statistische Werte festliegen, können andere Items eines neuen Tests auf einer gemeinsamen Schwierigkeitsskala lokalisiert (d. h. geeicht) werden.

Anweisungen Instruktionen zur Bearbeitung der Testaufgaben.

Aufgabe Teil eines Tests, komplexer als ein einzelnes Testitem. Eine Aufgabe bezieht sich im Allgemeinen auf eine mündliche oder schriftliche Leistung oder eine Reihe von in bestimmter Weise in Verbindung stehenden Items, z. B. einen Text zum Leseverstehen mit mehreren Multiple-Choice Aufgaben, die alle durch dieselbe Anweisung gesteuert werden.

Aufsichtsperson Person, die bei der Durchführung des Tests die Aufsicht im Prüfungsraum führt.

Aufgabenvorrevision Eine Phase in der Testproduktion, in der die Testentwickler das von den Testautoren eingereichte Material beurteilen und entscheiden, ob es abgelehnt werden soll, da es nicht den Testspezifikationen entspricht, oder ob es für die nächste Phase der Redaktion angenommen werden kann.

Auswerter Eine Person, die den Lösungen eines Teilnehmenden einen Zahlenwert zuordnet. Grundlage hierfür kann sowohl eine Experteneinschätzung sein als auch die weitgehend mechanische Verwendung eines Lösungsschlüssels.

Auswertung Zuordnung eines Wertes zu den Lösungen in einem Test. Dies bezieht sich sowohl auf die Expertenbeurteilung als auch die Verwendung eines Lösungsschlüssels, in dem alle akzeptablen Antworten aufgelistet sind.

Authentizität Zur Charakterisierung eines Tests bezeichnet der Begriff das Ausmaß, in dem der Test Sprachverwendung außerhalb der Prüfungssituation widerspiegelt, siehe auch **Testzweckmäßigkeit.**

Benotung Prozess der Umrechnung von Test- oder Punktwerten in Noten.

Beteiligte *(stakeholder)* Personen und Institutionen mit Interesse an dem Test, z. B. Prüfungsteilnehmerinnen und -teilnehmer, Schulen, Eltern, Arbeitgeber, Regierungen, Angestellte des Testanbieters.

Bewerter Eine Person, die einer Teilnehmerleistung in einem Test einen bestimmten Punktwert zuweist, wenn eine subjektive Bewertung erforderlich ist. Bewerterinnen und Bewerter sind normalerweise im entsprechenden Tätigkeitsbereich qualifiziert und müssen sich einem Qualifizierungsprozess unterziehen, der auch Kalibrierungen umfasst. Bei mündlichen Prüfungen haben Bewerter und Fragesteller, auch bezeichnet als Prüferinnen und Prüfer, etwas unterschiedliche Funktionen.

Bewertung Bewertung einer Leistung anhand einer Bewertungsskala. Die Bewertung wird durch qualifizierte Bewerterinnen und Bewerter vorgenommen.

Bewertungsskala Aus mehreren rangmäßig abgestuften Kategorien bestehende Skala, die für die subjektive Bewertung des freien schriftlichen und mündlichen Ausdrucks verwendet wird. Bei Sprachtests werden die Skalen üblicherweise mit Leistungsbeschreibungen versehen, wodurch die Interpretation erleichtert wird.

Deskriptor Eine kurze Beschreibung, die einem Wertebereich auf einer Bewertungsskala beigefügt wird und das Leistungsniveau derjenigen Personen kennzeichnet, die den entsprechenden Punktwert erzielen.

dichotomes Item Ein Item, das mit „falsch" oder „richtig" ausgewertet wird. Dichotome Items können beispielsweise *Multiple-Choice*-Aufgaben, Richtig/Falsch-Aufgaben oder Kurzantwortaufgaben sein.

Distraktor „Ablenker", das heißt eine nicht zutreffende Antwortmöglichkeit bei einem *Multiple-Choice*-Test.

Domäne (Sprachverwendungsbereich) Bereich des gesellschaftlichen Lebens, z. B. Ausbildung oder persönlicher Bereich, der definiert wird, um den zu prüfenden Inhalt oder die zu überprüfenden Fertigkeiten für einen Test zu beschreiben.

Doppelbewertung Beurteilungsmethode, bei der zwei Bewerter eine Testleistung unabhängig voneinander bewerten.

Echttest Ein Test, der zur Verwendung bereit steht und aus diesem Grunde vertraulich gehandhabt werden muss.

Eigenschaft Physisches oder psychisches Merkmal einer Person (wie die sprachliche Fähigkeit).

Erprobung Stadium im Prozess der Testentwicklung, in dem Aufgaben an einer repräsentativen Stichprobe aus der Zielpopulation erprobt werden, um die Schwierigkeit der Aufgaben zu ermitteln. Nach der Durchführung statistischer Analysen kann entschieden werden, welche Aufgaben in den endgültigen Test übernommen werden.

geschlossene Aufgabe Aufgabe, bei der die Antwort nicht selbstständig formuliert wird, das heißt, bei der keine produktive Sprachleistung gefordert ist. Beispiele: Richtig/Falsch-Aufgaben, *Multiple-Choice*- und Zuordnungsaufgaben.

Gewichtung Zuweisung unterschiedlicher maximal erreichbarer Punktwerte zu einem Testitem, einer Aufgabe oder einer Prüfungskomponente mit dem Ziel, den relativen Beitrag dieses Teils im Verhältnis zu den anderen Teilen am Zustandekommen eines Gesamtwertes zu verändern. Wenn beispielsweise für alle Items von Aufgabe 1 die doppelte Punktzahl vergeben wird, erhält Aufgabe 1 ein größeres Gewicht am Gesamtwert als die übrigen Aufgaben.

handlungsorientierter Ansatz Art, über Sprachfähigkeit nachzudenken, wobei Sprache als Werkzeug zur Durchführung kommunikativer Handlungen in einem gesellschaftlichen Kontext gesehen wird.

***High stakes*-Test** Hier geht es um das Maß, in dem das Prüfungsergebnis die Zukunft der Prüfungsteilnehmerinnen und -teilnehmer beeinflusst. Auswirkungen von *High stakes*-Tests sind gravierend für die Zukunft der Teilnehmenden.

Interaktivität Das Ausmaß, in welchem die Items und Aufgaben mentale Prozesse und Strategien erfordern, die auch bei Sprachhandlungen im tatsächlichen Leben notwendig wären. Siehe auch **Zweckmäßigkeit**.

Intervallskala Messskala, bei der der Abstand zwischen jedem Paar nebeneinanderliegender Messwerte gleich groß ist, bei der es allerdings keinen echten Nullwert gibt.

Item Jedes Einzelelement eines Tests, das getrennt bewertet wird. Beispiele sind: Lücken in einem *Cloze*-Test, *Multiple-Choice*-Frage mit drei oder vier Auswahlantworten, ein Satz, der grammatikalisch umzuformulieren ist, oder eine Frage, auf die ein Satz als Antwort erwartet wird.

Itemanalyse Analyse der Lösungen einzelner Testaufgaben, wobei üblicherweise die klassischen statistischen Kennwerte „Schwierigkeit" und „Trennschärfe" verwendet werden. Software wie MicroCAT wird dafür benutzt.

Itemdatenbank (Itembank) Werkzeug zur Verwaltung von Testaufgaben, das es durch Speicherung von Informationen über die Aufgaben ermöglicht, Tests mit beliebig festlegbarem Inhalt zu konstruieren.

kalibrieren (eichen) Im Probabilistischen Testmodell die Einschätzung der Schwierigkeit einer Reihe von Testitems.

Kalibrierung (Eichung) Der Prozess der Skalenbestimmung eines Tests. Eichung kann auf Ankeraufgaben aus unterschiedlichen Tests basieren mit dem Ziel, eine gemeinsame Schwierigkeitsskala (Theta-Skala) zu bilden. Wenn ein Test aus geeichten Aufgaben konstruiert wird, sind die Punktwerte des Tests direkter Ausdruck der Fähigkeit der Teilnehmenden, ausgedrückt durch die Lokalisierung der Person auf der Theta-Skala.

Konstrukt Eine angenommene Fähigkeit oder **Eigenschaft**, die nicht direkt gemessen oder beobachtet werden kann, z. B. bei Sprachprüfungen das Hörverstehen.

Korrelation Die Beziehung zwischen zwei oder mehr Messreihen, wobei geprüft wird, in welchem Ausmaß diese in der gleichen Weise variieren. Wenn beispielsweise Prüfungsteilnehmerinnen und -teilnehmer meist ähnliche Bewertungen bei zwei verschiedenen Tests erhalten, so ergibt sich eine positive Korrelation zwischen den Werten der beiden Tests.

Lösung (Antwort) Durch einen Test verursachtes Verhalten eines Prüfungsteilnehmers. Beispielsweise die gewählte Option bei einer *Multiple-Choice*-Aufgabe oder das Ergebnis eines schriftlich zu bearbeitenden Tests.

Lösungsschlüssel Eine Liste der akzeptablen Antworten auf die in einer Prüfung oder in einem Test gestellten Aufgaben. Ein Lösungsschlüssel ermöglicht einem Auswerter die genaue Zuweisung von Punktwerten in einem Test.

Logit Eine Maßeinheit, die in der Probabilistischen Testtheorie (Rasch-Modell) und in der Multifacetten-Rasch-Analyse benutzt wird.

manuelle Auswertung Bewertungsmethode, bei der die Bewerter über kein besonderes Fachwissen und subjektives Urteilsvermögen verfügen müssen. Sie bewerten nach einem Lösungsschlüssel, in dem alle zulässigen Lösungen für jede Aufgabe des Tests festgelegt sind.

Messskala Eine Skala mit Werten, die zur Messung des Unterschieds zwischen Prüfungsteilnehmenden, Items, Grenzwerten etc. hinsichtlich des Testkonstrukts benutzt wird. Eine Schwierigkeitsskala wird aufgebaut, indem statistische Verfahren auf die Teilnehmer-Lösungen und auf die Items angewendet werden (siehe Anhang VII). Schwierigkeitsskalen generieren mehr Informationen als Rohwerte, da sie nicht nur beispielsweise zeigen, welche Teilnehmer besser sind als andere, sondern auch, um wie viel jemand besser ist als ein anderer. Nominal- und Ordinalskalen werden oft als Schwierigkeitsskalen angesehen; diese Definition wurde in dem Handbuch jedoch nicht verwendet.

Mittelwert Ein Maß für zentrale Tendenz, oftmals auch als Durchschnitt bezeichnet. Der Mittelwert der Ergebnisse für eine Testdurchführung wird errechnet, indem alle Testwerte addiert und durch die Anzahl der Testwerte geteilt werden.

Modell der Sprachverwendung Eine Beschreibung aller für die Sprachverwendung benötigten Fähigkeiten und Kompetenzen und der Art, wie sie miteinander in Verbindung stehen. Dieses Modell ist die Grundlage für die Testentwicklung.

Modellanpassung Wird ein Modell (wie das Rasch-Modell) in der statistischen Analyse verwendet, muss beachtet werden, wie gut die Daten und das Modell zusammenpassen. Ein Modell ist ein Ideal, wie die Daten sein sollten; eine perfekte Modellanpassung ist also nicht vorgesehen. Ein hohes Maß an Nichtanpassung kann jedoch bedeuten, dass die Rückschlüsse über die Daten falsch sind.

Multifacetten-Rasch-Modell (polytomes Rasch-Modell) Eine Erweiterung des Rasch-Modells, die die Abbildung von Antwortwahrscheinlichkeiten auf der Grundlage der Kombination zusätzlicher Parameter gestattet. So kann die Bearbeitung einer schriftlichen Aufgabe so abgebildet werden, dass sich darin zusätzlich zur Aufgabenschwierigkeit auch die Strenge des Bewerters abbildet. Das polytome Rasch-Modell kann beispielsweise mit dem Programm FACETS gerechnet werden.

Note Ein Prüfungsergebnis kann auf einer Notenskala dargestellt werden, die beispielsweise von 1 bis 6 reicht, wobei 1 die beste Note ist, 2 mit guten, 3 mit befriedigenden, 4 mit ausreichenden Leistungen bestanden sowie 5 und 6 durchgefallen bedeutet.

objektive Auswertung Objektiv ausgewertet werden Items, die mit einem Lösungsschlüssel ausgewertet werden, ohne dass die Meinung von Experten oder eine subjektive Beurteilung in die Auswertung einfließt.

offene Aufgabe Aufgabentyp in einem schriftlichen Test, bei dem die Teilnehmenden eine Lösung frei ergänzen müssen und diese nicht aus mehreren vorgegebenen auswählen können. Diese Art von Aufgabe soll zu möglichst freien Antworten führen, deren Umfang zwischen einzelnen Wörtern und einem kompletten Aufsatz liegen kann. Der Lösungsschlüssel sieht deshalb ggf. eine Reihe akzeptabler Lösungen vor.

optischer Belegleser (OMR-Technologie) Ein elektronisches Gerät, das Informationen direkt von einem Antwortbogen abliest. Test- oder Prüfungsteilnehmer können ihre Lösungen auf einem Antwortbogen markieren. Diese Informationen werden dann direkt in den Computer eingelesen. Wird auch als Scanner bezeichnet.

***Partial-Credit*-Item** Item, das so bewertet wird, dass auch teilweise richtige Antworten Punkte erhalten. Zum Beispiel können die Punkte auf eine Antwort 0, 1, 2 oder 3 sein, je nach Grad der Korrektheit, so wie im Lösungsschlüssel beschrieben.

Pilottest Erprobung des Testmaterials auf sehr kleiner Stufe, zum Beispiel, indem Kollegen gebeten werden, die Items zu lösen und zu kommentieren.

Praktikabilität Das Maß der Möglichkeiten, einen Test so zu entwickeln, dass er mit den gegebenen Ressourcen den Anforderungen entspricht.

probabilistische Testmodelle Eine Gruppe mathematischer Modelle zur Herstellung von Beziehungen zwischen individuellen Testleistungen und dem Fähigkeitsniveau einer Person. Diese Modelle basieren auf der grundlegenden theoretischen Annahme, wonach der zu erwartende Erfolg bei der Bearbeitung einer Aufgabe eine Funktion sowohl des Schwierigkeitsgrades der Aufgabe als auch des individuellen Fähigkeitsniveaus der jeweiligen Person ist. Die Modelle werden auch unter dem Begriff „Item-Response-Theorie“ zusammengefasst.

Prüfungstermin	Datum oder Zeitraum, an oder in dem eine Prüfung stattfindet. Viele Prüfungen werden zu bestimmten festgelegten Terminen mehrfach im Jahr durchgeführt, während andere nach Bedarf angeboten werden.
Rasch-Modell	Mathematisches Modell, auch als einfaches logistisches Modell bezeichnet, das eine Beziehung zwischen der Wahrscheinlichkeit, dass eine Person eine Aufgabe löst, und der Differenz zwischen der Fähigkeit dieser Person und der Schwierigkeit der Aufgaben annimmt. Mathematisch entspricht das Rasch-Modell der Item-Response-Theorie.
Register	Klar abgegrenzte Merkmale der gesprochenen oder geschriebenen Sprache bezogen auf ein bestimmtes Handlungsfeld oder einen bestimmten Grad der Förmlichkeit und Höflichkeit.
Reliabilität (Zuverlässigkeit)	Konsistenz oder Stabilität der mit einem Test vorgenommenen Messungen. Je reliabler ein Test ist, desto geringer ist der enthaltene Messfehler. Ein Test, der einen systematischen Fehler enthält, z. B. eine bestimmte Personengruppe begünstigt oder benachteiligt, kann reliabel sein, nicht aber valide.
Rohwert	Testwert, der noch nicht statistisch durch eine Transformation, Gewichtung oder Skalierung verändert wurde.
Schwierigkeitswert *(facility)*	Schwierigkeit einer Aufgabe, ausgedrückt durch den Anteil richtiger Lösungen auf einer von 0 bis 1 reichenden Skala, mitunter auch in Prozentwerten ausgedrückt.
Skala	Satz von Zahlen oder Kategorien, die verwendet werden, um etwas zu messen. Man unterscheidet zwischen vier Skalenniveaus: Nominal-, Ordinal-, Intervall- und Verhältnisskala.
Spannweite	In der Statistik Maß für die Verteilung von Beobachtungswerten. Die Spannweite ist die Differenz zwischen dem höchsten und dem niedrigsten beobachteten Wert.
Standard-abweichung	Maß für die Streuung von Beobachtungen um den arithmetischen Mittelwert herum. Die Standardabweichung ist die Quadratwurzel der Varianz. Bei Vorliegen einer Normalverteilung liegen 68 % der Fälle der Stichprobe innerhalb von einer Standardabweichung und 95 % innerhalb von zwei Standardabweichungen rechts und links vom Mittelwert.
Standardsetzung	Prozess der Festlegung von Grenzwerten beim Test (z. B. die Grenze zwischen bestehen und nicht bestehen) und somit die Festlegung der Bedeutung von Prüfungsergebnissen.
subjektive Bewertung	Subjektiv bewertet werden Items, die unter Zuhilfenahme von Expertenmeinungen oder durch subjektive Beurteilung bewertet werden.
Testentwickler	Jemand, der sich mit der Entwicklung neuer Tests beschäftigt.
Testerstellung	Prozess der Auswahl von Texten, das Schreiben der Aufgaben und die Zusammenstellung zu einem Test. Dem geht oftmals eine Vorerprobung oder eine Erprobung von Aufgabenmaterialien voraus. Aufgaben können auch aus Aufgabenbanken entnommen werden.
Testspezifikationen	Eine Beschreibung der Merkmale der Prüfung, z. B. was geprüft wird, wie es geprüft wird, Angaben zur Zahl und Länge der Aufgaben, Aufgabentypen etc.

Testteil	Teil eines Tests, der oft als unabhängiger Test vorgegeben wird, mit eigenem Aufgabenheft und eigener Zeitbegrenzung. Testteile beziehen sich oft auf spezielle Einzelfertigkeiten und sind überschrieben mit Titeln wie „Hörverstehen“ oder „Aufsatz“.
Testversionen (Parallelformen)	Verschiedene Versionen desselben Tests werden als äquivalent angesehen, weil sie in der gleichen Weise spezifiziert sind und das gleiche Merkmal messen. Um die strengen Äquivalenzbedingungen der klassischen Testtheorie zu erfüllen, müssen die verschiedenen Testversionen die gleiche mittlere Schwierigkeit, Varianz und Kovarianz aufweisen, wenn sie von denselben Personen bearbeitet werden. In der Praxis ist es schwierig, echte Parallelformen zu entwickeln.
Textbasiertes Item	Item, das auf einem Teil eines zusammenhängenden Textes basiert, z. B. *Multiple-Choice*-Aufgabe, die sich auf einen Lesetext bezieht.
Trennschärfe	Die Fähigkeit einer Aufgabe, zwischen stärkeren und schwächeren Prüfungsteilnehmenden zu unterscheiden. Es gibt unterschiedliche Kennwerte dafür, siehe Anhang VII.
unabhängiges Item (diskretes Item)	Ein in sich abgeschlossenes Item, das nicht mit anderen Items oder sonstigem Material verbunden ist.
Validierung	Prozess des Nachweises der Validität der Interpretation von Prüfungsergebnissen, wie sie vom Prüfungsanbieter empfohlen wird.
Validität	Ausmaß, in dem Testwerte angemessene, sinnvolle und nützliche Schlussfolgerungen gemäß Messintention des Tests zulassen.
Validitätsbeleg (Validitäts-argumentation)	Eine umfangreiche Reihe von Aussagen und diese unterstützenden Belegen, die die Validität der empfohlenen Interpretation von Testergebnissen begründen soll.
Verlinkung (Anbindung)	Der Prozess, der die Ergebnisse eines Tests oder eines Testformats „übersetzt“, so dass sie in Bezug auf die Ergebnisse eines anderen Tests oder eines anderen Testformats verstanden werden können. Dieses Verfahren trägt dazu bei, Unterschiede bei der Testschwierigkeit oder den Fähigkeiten der Teilnehmenden auszugleichen.
Verzerrung *(bias)*	Begünstigung oder Benachteiligung eines bestimmten Anteils der Teilnehmer-Population, hervorgerufen durch bestimmte Merkmale des Tests oder der Aufgabe, die nicht Gegenstand dessen sind, was gemessen werden soll.
Verwendung des Tests	Die Argumentation zur Testverwendung ist der Teil der Validitätsargumentation, der begründet, wie das Testergebnis interpretiert werden darf.
Vorerprobung *(trialling)*	Stadium der Entwicklung von Testaufgaben, in dem geprüft wird, ob der Test in der erwarteten Weise funktioniert. Es handelt sich dabei oft um offene Aufgaben, wie z. B. einen Aufsatz, die von einer kleineren Probandengruppe bearbeitet werden.
Vorgabe *(prompt)*	Bezeichnet alle Materialien (graphisches Material oder Texte), die in Tests zum Sprechen oder Schreiben eingesetzt werden, um Prüfungsteilnehmerinnen und -teilnehmer zu schriftlichen oder mündlichen Äußerungen anzuregen.
Vorgabe *(input)*	Material in einer Testaufgabe, auf das die Teilnehmenden in angemessener Weise reagieren sollen. Zum Beispiel kann in einem Hörverstehenstest die Vorgabe in einem auf Band aufgenommenen Text bestehen und einigen dazugehörigen schriftlichen Aufgaben.

Wirkung Auswirkung einer Prüfung sowohl auf die Gesellschaft und auf Bildungsprozesse als auch auf einzelne Personen, die von dem Prüfungsergebnis betroffen sind.

Zuordnungsaufgabe Aufgabentyp, der eine Zuordnung der Elemente zweier getrennter Listen fordert. Eine Art von Zuordnungsaufgabe besteht darin, dass der richtige Begriff zur Vervollständigung eines Satzes auszuwählen ist. Eine Zuordnungsaufgabe zu einem Lesetext könnte darin bestehen, dass aus einer Liste etwas wie eine Freizeitbeschäftigung oder ein Buch auszuwählen ist, das zu einer Person passen soll, deren spezielle Bedürfnisse beschrieben werden.

Zweckmäßigkeit Der Begriff Zweckmäßigkeit in Bezug auf Tests (Bachman und Palmer 1996) bezeichnet die Vorstellung, dass ein Test dann am sinnvollsten ist, wenn ein Gleichgewicht zwischen Validität, Reliabilität, Authentizität, Interaktion, Wirkung und Praktikabilität besteht.

Dieses Glossar wurde aus dem *Multilingual Glossary of Language Testing Terms*, erstellt von ALTE (ALTE Members 1998), und dem *Dictionary of Language Testing* (Davies et al. 1999) zusammengestellt, beide von Cambridge University Press in der Serie *Studies in Language Testing* veröffentlicht. Weitere Einträge wurden nach Bedarf hinzugefügt.

Danksagung

Dieses Handbuch ist eine überarbeitete Neuauflage eines früheren Handbuchs, das 2002 vom Europarat unter dem Titel *Language Examination and Test Development* herausgegeben wurde. Hierbei handelte es sich um eine Version des 1996 von ALTE für den Europarat erstellen Dokuments mit dem Titel *Users' Guide for Examiners.*

Der Europarat bedankt sich für die Unterstützung durch:

die Association of Language Testers in Europe (ALTE) für die Überarbeitung des Dokuments

bei der Redaktionsgruppe für diese Überarbeitung:

David Corkill
Michael Corrigan
Neil Jones
Michael Milanovic
Martin Nuttall
Nick Saville

bei den Mitgliedern der „ALTE CEFR/Manual Special Interest Group" und ihren Kollegen für die Bereitstellung zusätzlichen Materials und für die mehrmalige Begutachtung der Textentwürfe:

Elena Archbold-Bacalis
Sharon Ashton
Andrew Balch
Hugh Bateman
Lyan Bekkers
Nick Beresford-Knox
Cris Betts
Margherita Bianchi
Inmaculada Borrego
Jasminka Buljan Culey
Cecilie Carlsen
Lucy Chambers
Denise Clarke
Maria Cuquejo
Emyr Davies
Desislava Dimtrova
Angela ffrench
Colin Finnerty
Anne Gallagher
Jon-Simon Gartzia
Annie Giannakopoulou
Begona Gonzales Rei
Guiliana Grigorova
Milena Grigorova
Ines Hälbig
Berit Halvorsen
Marita Harmala
Martina Hulešová
Nurita Jornet
Marion Kavallieros
Gabriele Kecker
Kevin Kempe
Wassilios Klein
Mara Kokina
Zsofia Korody
Henk Kuijper
Gad Lim
Juvana Llorian
Karen Lund
Lucia Luyten
Hugh Moss
Tatiana Nesterova
Desmond Nicholson
Gitte Østergaard Nielsen
Irene Paplouca
Szilvia Papp
Francesca Patrizzi
Jose Ramón Rarrondo
Jose Pascoal
Roberto Perez
Michaela Perlmann-Balme
Tatiana Perova
Sibylle Plassmann
Laura Puigdomenech
Meilute Ramoniene
Lýdia Rihová
Shelagh Rixon
Martin Robinson
Lorenzo Rocca
Shalini Roppe
Dittany Rose
Angeliki Salamoura
Lisbeth Salomonsen
Gergio Silfer
Gabriela Šnaidaufová
Ioana Sonea
Annika Spolin
Stefanie Steiner
Michaela Stoffers
Gunlog Sundberg
Lynda Taylor
Julia Todorinova
Rønnaug Katharina Totland
Gerald Tucker
Piet van Avermaet
Mart van der Zanden
Elorza Juliet Wilson
Beate Zeidler
Ron Zeronis

bei den Begutachtern des Europarats:

Neus Figueras
Brian North
Johanna Panthier
Sauli Takala

bei dem Verlegerteam:

Rachel Rudge
Gary White

Die vorliegende Übersetzung wurde angefertigt von:

Sandra Hohmann
Heinrich Rübeling
Beate Zeidler
Gudrun Klein
Viola Stübner
Sibylle Plassmann
Jaqueline Thommes

Die Vereinigung von Sprachtestanbietern in Europa (*Association of Language Testers in Europe* – ALTE) ist eine Internationale Nicht-Regierungsorganisation mit Beraterstatus im Europarat. Sie hat dazu beigetragen, das vom Europarat veröffentlichte Instrumentarium zur Ergänzung des GER zu erstellen. Neben dem vorliegenden Handbuch hat ALTE auch das *EAQUALS/ALTE European Language Portfolio* (ELP) und die GER-Raster zur Analyse mündlicher und schriftlicher Aufgaben entwickelt.

Gemeinsam mit der Abteilung für Sprachenpolitik des Europarates will ALTE die Nutzer dieses Instrumentariums dazu anregen, den GER in ihrem eigenen Kontext zu nutzen, um ihre Ziele erfolgreich umzusetzen.

Das vorliegende Handbuch wurde von ALTE in englischer Sprache erstellt und unter Federführung der gemeinnützigen telc GmbH ins Deutsche übersetzt.

Im Auftrag des Europarats übersetzt von:

telc GmbH

Bleichstraße 1
60313 Frankfurt am Main
Deutschland

www.telc.net