Erstellt von ALTE
im Auftrag des Europarats/Abteilung für Sprachenpolitik

Handbuch zur Entwicklung und Durchführung von Sprachtests

Zur Verwendung mit dem GER

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

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.

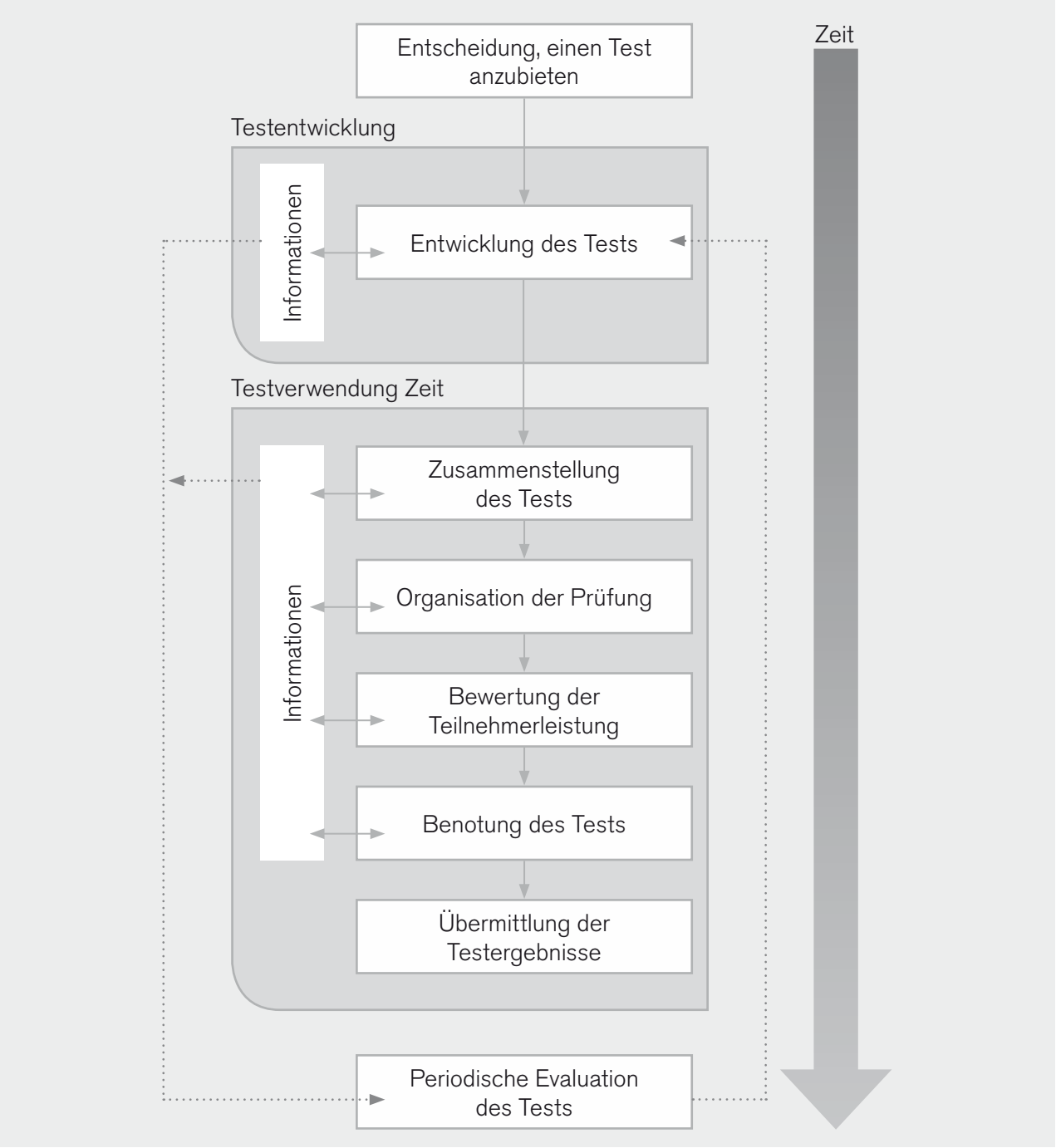

Abb. 15: Der allgemeine Testzyklus und die periodische Evaluation

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).

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.

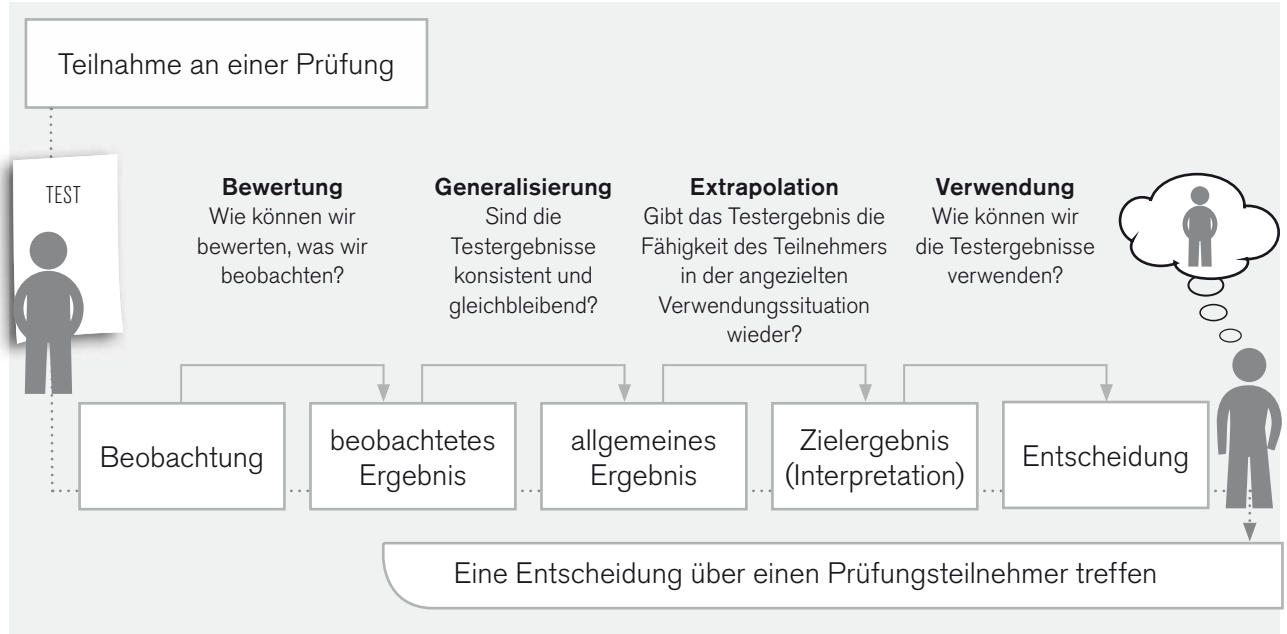

Abb. 16: Logische Folge in einer Validitätsargumentation (angepasst von Cane, Crooks, Cohen 1999, Bachman 2005)

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.

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