GOETHE-ZERTIFIKAT A2

STANDARD SETTING UND BENCHMARKING

ERGEBNISSE

München, 18.–19.01.2016

www.goethe.de

GOETHE INSTITUT

Sprache. Kultur. Deutschland.

IMPRESSUM

2016 Goethe-Institut, Zentrale, Bereich 41, München

Gesamtkoordination
Michaela Perlmann-Balme, Goethe-Institut, Abteilung Sprache, Zentrale, München

Autorinnen
Doris Hennemann
Michaela Perlmann-Balme
Claudia Stelter

Datenanalyse
Jane Lloyd, Cambridge English Language Assessment

Redaktion
Falko Röhrs
Beate Schnorfeil

Coverfoto
Goethe-Institut/Valentin Fanel Badiu

Gestaltung
Falko Röhrs

Material
Erwachsene: Items Lesen, Hören, Aufgaben Schreiben und Sprechen
https://www.goethe.de/de/spr/kup/prf/prf/gzsd2/ub2.html

Jugendliche: Items Lesen, Hören, Aufgaben Schreiben und Sprechen
https://www.goethe.de/de/spr/kup/prf/prf/gzfit2/uf2.html

Das Werk und seine Teile sind urheberrechtlich geschützt.
© 2016 Goethe-Institut

Inhalt

Vorwort

Das Goethe-Institut nimmt in regelmäßigen Abständen eine Revision seiner Deutschprüfungen vor, um sie an die gesellschaftliche Entwicklung in den deutschsprachigen Ländern und an den Stand der Testforschung anzupassen. Bei Revisionen wird außerdem das Ziel verfolgt, die Prüfung möglichst genau auf den *Gemeinsamen europäischen Referenzrahmen für Sprachen* (GeR) und dessen Niveaubeschreibungen zu beziehen.

Inzwischen spielt die Einhaltung prüfungsrelevanter Qualitätsstandards (vgl. die Richtlinien der European Association for Language Testing and Assessment (EALTA) von 2006 und die Minimalstandards der ALTE von 2007) eine zunehmend wichtige Rolle. Die Association of Language Testers in Europe (ALTE) fordert dazu in ihren Mindeststandards: „Wenn Ihre Prüfung sich auf ein externes Referenzsystem bezieht (z. B. den *Gemeinsamen europäischen Referenzrahmen),* stellen Sie sicher, dass Sie diesen Bezug durch ein angemessenes methodisches Vorgehen nachweisen." (ALTE 2007) Zu dem hier geforderten methodischen Vorgehen gehört das sog. Standard Setting und Benchmarking mit externen Experten. Dabei geht es um den Nachweis, dass die Prüfungsanforderungen und die erhobenen Teilnehmerleistungen mit der Definition des angestrebten Niveaus im *Referenzrahmen* kompatibel sind. Zum zweiten geht es um die Feststellung bzw. Bestätigung der geplanten Bestehensgrenze (cut-off). Welches Ergebnis muss erreicht werden, um die Prüfung zu bestehen? Welche Leistungen reichen nicht aus? Bestehensgrenzen sind für das Prüfen und Testen zentral, weil darauf gestützte Entscheidungen für Testteilnehmende ebenso wie für Entscheidungsträger, beispielsweise für eine Schule, für einen Kursanbieter oder für eine Behörde, folgenreich sein können (Glaboniat/Perlmann-Balme/Studer 2013: 73).

Am 18. und 19. Januar 2016 fand in der Zentrale des Goethe-Instituts in München die Konferenz „Standard Setting und Benchmarking" zum neu entwickelten *Goethe-Zertifikat A2* statt. Diese neue Prüfung löst im Portfolio des Goethe-Instituts die bisherigen Prüfungen *Goethe-Zertifikat A2: Start Deutsch 2* und *Goethe-Zertifikat A2: Fit in Deutsch 2* zum 01. April 2016 ab. Die sogenannten Modellsätze zu diesen beiden Versionen finden sich als PDF zum Herunterladen auf der Homepage des Goethe-Instituts (siehe Seite Impressum).

Das Programm wurde auf der Grundlage der im Handbuch *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching,*

Assessment (2009) des Europarats vorgeschlagenen Schritte durchgeführt. An der Konferenz nahmen ausgewiesene Expertinnen und Experten aus dem Bereich Deutsch als Fremdsprache und Sprachlehrende sowie aus Bildungsinstitutionen teil. Insgesamt waren 37 Expert/-innen aus sechs europäischen Ländern anwesend.

Vertreten waren folgende Institutionen:

Testentwickler, pädagogische und testmethodische Einrichtungen:

Association of Language Testers in Europe (ALTE), Cambridge English Language Assessment, Centraal Instituut voor Toetsontwikkeling Niederlande (CITO), Goethe-Institut e. V., ÖSD, TestDaF-Institut, Universität Freiburg/Schweiz Bereich Mehrsprachigkeitsforschung und Fremdsprachendidaktik

Hochschulen, Universitäten:

Freie Universität Bozen, Friedrich-Schiller-Universität Jena, Jagellionen-Universität Krakau, Ludwig-Maximilians-Universität München, Universität Freiburg/Schweiz, Zürcher Hochschule für Angewandte Wissenschaften

Verlage für Deutsch als Fremdsprache:

Cornelsen, Hueber, Klett, Klett Hellas, Klett-Langenscheidt, Spotlight

Landesverbände der Volkshochschulen:

Bayern, Saarland, Sachsen-Anhalt

Anbieter von Deutschkursen und Prüfungszentren:

did deutsch-institut, Goethe-Institut Berlin, Düsseldorf, Freiburg, Mannheim, Münchner Volkshochschule, Sprachen und Dolmetscher Institut München, Sprachenzentrum der Universität des Saarlandes

An dieser Stelle danken wir noch einmal herzlich allen Expertinnen und Experten für ihre Zeit und das hohe Engagement.

Die Erfahrungen, die wir mit dem Standard Setting und Benchmarking gemacht haben, veröffentlichen wir in Form dieses Berichts, um Vorgehensweisen und Ergebnisse offenzulegen. Zweifellos enthält er auch diskutable Passagen, was bei urteilsbasierten Daten in der Natur der Sache liegt. Wir verstehen diesen Bericht als Beitrag zu einem empirisch gestützten Validitätsargument (Bachman, Palmer 2010; vgl. Weir 2005).

Die Autorinnen

1 Das Projekt Revision des *Goethe-Zertifikats A2*

Das *Goethe-Zertifikat A2* wurde vom Goethe-Institut, Zentrale, Bereich 41, entwickelt. Dabei gab es folgende Meilensteine:

Bedarfserhebung: Eine im Frühjahr 2013 weltweit an Prüfungszentren des Goethe-Instituts durchgeführte Bedarfs- und Zielgruppenanalyse bildete die Grundlage für die Konzeption der neuen Prüfung.

Entwurf des Testkonstrukts: Unter Hinzuziehung ausgewiesener Expertinnen und Experten begann die Entwicklung des *Goethe-Zertifikats A2* mit der Definition des Prüfungskonstrukts. Um die neue Sprachprüfung auf dem Sprachniveau A2 des *Referenzrahmens* zu positionieren, wurde ein zweistufiges Verfahren aus qualitativer Begutachtung durch externe Testexpert/-innen sowie aus einer Erprobung der Entwurfsfassung angewendet.

Externe Begutachtung: Zunächst wurde das Expertenurteil von fünf Gutachtern zum Testkonstrukt eingeholt. Mit einer detaillierten Beschreibung der Prüfungsziele und -inhalte wurde ein transparenter Bezug zu den Kann-Beschreibungen des *Referenzrahmens* für das Niveau A2 sichergestellt. Dieser Bezug wurde durch Gutachten externer Expertinnen und Experten bestätigt.

Erprobung des Entwurfs: Durchgeführt wurde eine weltweite Validierung des Testmodells, die u. a. zum Ziel hatte, die Brauchbarkeit der Aufgabentypen zu überprüfen. Hierbei wurde neben der Frage des Schwierigkeitsgrades besonders auf die Akzeptanz und Praktikabilität der Aufgaben, den Zeitbedarf und die Länge der produzierten Texte geachtet. Die Ergebnisse der Erprobung wurden statistisch analysiert. Durch diese Analysen ließen sich Schlüsse ziehen auf die Performanz sowie auf Schwierigkeit und Trennschärfe der erprobten Testentwürfe.

Kommunikation der Testmaterialien: Im August 2014 wurden die überarbeiteten Testentwürfe als *Modellsatz Erwachsene* und *Modellsatz Jugendliche* auf der Internetseite des Goethe-Instituts als vorläufige Version veröffentlicht. Es folgte im Oktober 2015 die Veröffentlichung des Handbuchs *Prüfungsziele, Testbeschreibung*. Es beschreibt alle Grundlagen der Testentwicklung und alle sprachlichen Inhalte inklusive einem Inventar mit

Listen zu Sprachhandlungen, Wortschatz und Strukturen (Hennemann et al. 2015). Trainingsmaterialien für die Ausbildung der Prüfenden und Bewertenden wurden intern publiziert, um damit Multiplikatorinnen und Multiplikatoren sowie alle Bewertenden weltweit online oder in Präsenzseminaren zu schulen.

Standard Setting und Benchmarking: Mit diesen Arbeitsschritten wird die Testentwicklung abgeschlossen. Die Ergebnisse, die im vorliegenden Bericht dokumentiert sind, fließen in die Fortschreibung der Prüfung ein. Der Terminus Standard Setting meint ein Bündel von strukturierten Verfahren, deren Ziel darin besteht, rezeptive Lernerleistungen auf verbal definierte Niveaustufen wie diejenigen des GeR zu beziehen (Kenyon 2013: 1). Dabei bilden menschliche Urteile in Form von individuellen und in der Gruppe ausgehandelten Entscheidungen (Kantarcioglu, Papageorgiou 2011: 99f) die Basis. Gegenstand von Standard Settings sind rezeptive Leistungen, im vorliegenden Fall Testresultate zum Hör- und Leseverstehen, die von einem Panel aus Fachleuten danach beurteilt werden, ob sie für das Erreichen einer Niveaustufe genügend und/oder typisch sind. Für die Beurteilung produktiver schriftlicher und mündlicher Testresultate beurteilt das Panel aus Fachleuten eine möglichst große Zahl an Kandidatenleistungen. Für diesen Arbeitsschritt hat sich der Begriff Benchmarking etabliert.

2 Arbeitsgruppen zum Standard Setting und Benchmarking

2.1 Lesen

Leitung: Doris Hennemann

Assistenz: Michaela Perlmann-Balme, Linda Fromme

Teilnehmende: Ulrike Arras, Freie Universität Bozen
Gudula Bieber-Reynartz, Münchner Volkshochschule
Dominik Breithaupt, SDI München
Kirsten Bröcker, Landesverband der VHS Sachsen-Anhalt
Silvia Demmig, Friedrich-Schiller-Universität Jena
Armin Göbels, Goethe-Institut Berlin
Dorrie Goossens, Cito Niederlande
Corinna Hilger, Cornelsen Schulverlage
Ina Hoischen, Goethe-Institut Zentrale, Bereich 42
Silke Jacobs, Goethe-Institut Düsseldorf
Tanja Krüger, Goethe-Institut Zentrale, Bereich 41
Stefan Laub, did deutsch-institut
Uta Loumiotis, Klett Hellas
Florian Nimmrichter, Österreichisches Sprachdiplom
Stefanie Plisch de Vega, Ernst Klett Sprachen
Annerose Remus, Klett-Langenscheidt Verlag
Irmingard Staudigel, Landesverband der VHS Bayern
Virginia Suter Reich, Zürcher Hochschule für Angewandte Wissenschaften
Elisabetta Terrasi-Haufe, Ludwig-Maximilians-Universität München

2.1.1 Verfahren

Ziel der Arbeitsgruppe war es, für den Prüfungsteil *Lesen* zu bestimmen, wie viele Aufgaben richtig gelöst werden müssen, um diesen zu bestehen. Dazu wurde in einem zweistufigen Verfahren ein Leistungsstandard (cut score) festgelegt. Mit dieser Festlegung sollte sichergestellt werden, dass die in Form von Aufgaben operationalisierten Anforderungen der Prüfung *Goethe-Zertifikat A2* (für Jugendliche und Erwachsene) in dem Prüfungsteil *Lesen* dem angezielten Niveau A2 des *Gemeinsamen europäischen Referenzrahmens für Sprachen (GeR)* entsprechen.

Bestimmt wurde die zu erreichende Grenze mit der Bookmark-Methode (Karantonis/Sireci: (2006; Kecker: 2010). Grundlage dieses Item Response Theory (IRT)-basierten Verfahrens ist eine Anordnung der Items in einem „Booklet“, nicht nach deren Abfolge in der Prüfung, sondern nach ihrem statistischen Schwierigkeitswert aufsteigend, beginnend mit dem leichtesten Item. Die Schwierigkeitswerte basieren auf einer Rasch-Analyse der Rückläufe aus Erprobungen des Modellsatzes (n=209 Modellsatz_Erwachsene und n=127 für Modellsatz_Jugendliche). Das Booklet enthält 20 Items des Prüfungsteils *Lesen.* Das Testkonstrukt für die Jugendversion und die Erwachsenenversion ist analog gestaltet. Die Items für das Standard Setting stammen sowohl aus dem Modellsatz für Erwachsene (Item 1 bis 5 und 11 bis 20) als auch aus dem für Jugendliche (Item 6 bis 10). Diese finden sich auf der Homepage des Goethe-Instituts (siehe Impressum). Jedes Item wurde mit Lösung auf einer separaten Seite dargestellt. Wenn mehrere Items zu einem Text gehörten, wurden sie ggf. auf nicht aufeinanderfolgenden Seiten präsentiert, wenn die Schwierigkeitswerte dies vorgaben.

Die Jurorinnen und Juroren hatten die Aufgabe, zu entscheiden, was ihrer Meinung nach eine knapp genügende A2-Leistung sei. Ihre Entscheidung sollten sie auf zwei Konzepte stützen: erstens auf das Konzept einer Person, die hinsichtlich des Niveaus A2 minimal kompetent ist, und zweitens auf das Konzept der Lösungswahrscheinlichkeit. Bei der minimal kompetenten Person (Zeidler 2016) mussten sich die Jurorinnen und Juroren eine/-n Prüfungsteilnehmende/-n mit einer Kompetenz am unteren Rand von A2 vorstellen. Bei der Lösungswahrscheinlichkeit mussten sie präzisieren, was es für sie bedeutet, ein Item mit relativ hoher Wahrscheinlichkeit korrekt lösen zu können. Die Jurorinnen und Juroren sollten sich vorstellen, dass die mindestkompetente Person das Item in zwei von drei Fällen richtig löst oder dass zwei von drei mindestkompetenten Personen das Item korrekt lösen. Auf dieser konzeptionellen Grundlage arbeiteten die Jurorinnen und Juroren die Item-Booklets Seite für Seite durch. Sie bewerteten die Schwierigkeit der Items aus der Sicht der mindestkompetenten A2-Person und entschieden, bei welchem Item die Wahrscheinlichkeit größer als Zweidrittel ist, dass diese Person das Item korrekt löst. Dieses Item musste durch Markieren der betreffenden Seite im Item-Booklet gekennzeichnet werden. Die Markierung steht für die Meinung der Jurorinnen und Juroren, dass die Items, die im Booklet auf den Seiten vor der markierten Seite stehen, von der mindestkompetenten Person mit einer Wahrscheinlichkeit 0,67 oder höher korrekt gelöst werden.

Die Arbeit vollzog sich in drei Phasen. In der ersten Phase stand die auch für Fachleute in Deutsch als Fremdsprache immer wieder notwendige Beschäftigung mit dem *Gemeinsamen europäischen Referenzrahmen für Sprachen (GeR)* im Mittelpunkt. Nachdem ein globales Vertrautmachen mit dem Niveau A2 bereits in der Gesamtgruppe vorgenommen worden war, konzentrierte sich die Gruppe *Lesen* auf die Deskriptoren mit Relevanz für diese Fertigkeit. Gearbeitet wurde mit einer Zuordnungsaufgabe, bei der es darum ging, das Niveau einer Reihe von Kann-Beschreibungen des *GeR* aus den aufgabenorientieren Skalen zum Lesen zu erkennen. Vorgelegt wurden Kann-Beschreibungen der Niveaus A1, A2, B1 ohne Niveau-Angabe. Das Erkennen des Niveaus sollte ausschließlich auf Basis von Niveauindikatoren in den Deskriptoren erfolgen. Als Hilfestellung dienten den Teilnehmenden die Beschreibungen der Niveaubereiche A1, A2, B1 aus den Skalen „Leseverstehen allgemein". Besonders fokussiert wurde bei dieser Arbeitsgruppenaktivität der Übergang von A1 zu A2.

Im Anschluss an die kurze Vorstellung des Test-Konstrukts *Lesen* und den Erläuterungen zur Umsetzung der Konstrukte in Aufgaben folgte eine zweite Phase des Vertrautmachens, bei der vom Europarat als auf A2-Niveau ausgewiesene Prüfungsaufgaben zur Diskussion gestellt wurden. Durch diese Diskussion in der Gruppe sollten sich die Teilnehmenden auf das A2-Niveau kalibrieren.

Als dritte Phase folgte das Standard Setting, bei dem das Item-Booklet *Lesen* zweimal durchgearbeitet wurde. In Runde 1 beurteilten die Jurorinnen und Juroren die Items in Einzelarbeit und setzten die Markierung im Item-Booklet. Sie gaben ihre Urteile anonym ab, d. h. jede/jeder hatte eine Identifikationsnummer. Die Ergebnisse dieser 1. Runde wurden registriert, als Säulendiagramme aufbereitet (vgl. Schaubild 1) und in dieser Form als Input für die Diskussion verwendet, die im Anschluss an Runde 1 stattfand. Diskutiert wurde in vier Teilgruppen von jeweils vier bis fünf Personen, wobei bei der Zusammensetzung der Diskussionsgruppen darauf geachtet wurde, Teilnehmende mit weiter auseinander liegenden Bookmarks zusammenzubringen. Ziel dieser Diskussionen war es, die Einzelvoten zu begründen. Es sollten Argumente für Entscheidungen ausgetauscht und insbesondere auch Gründe für stärker divergierende Voten vorgebracht und verglichen werden.

Nach der Diskussion in Teilgruppen setzten die Jurorinnen und Juroren in Runde 2 wieder individuell ihre Markierung im Item-Booklet. Dabei stand es ihnen frei, ihre Markierung aus der 1. Runde zu übernehmen oder diese unter dem Eindruck der Diskussion neu zu setzen. Die Ergebnisse der 2. Runde wurden ebenfalls registriert, aufbereitet und präsentiert (vgl. Schaubild 2).

2.1.2 Ergebnisse

Schaubild 1

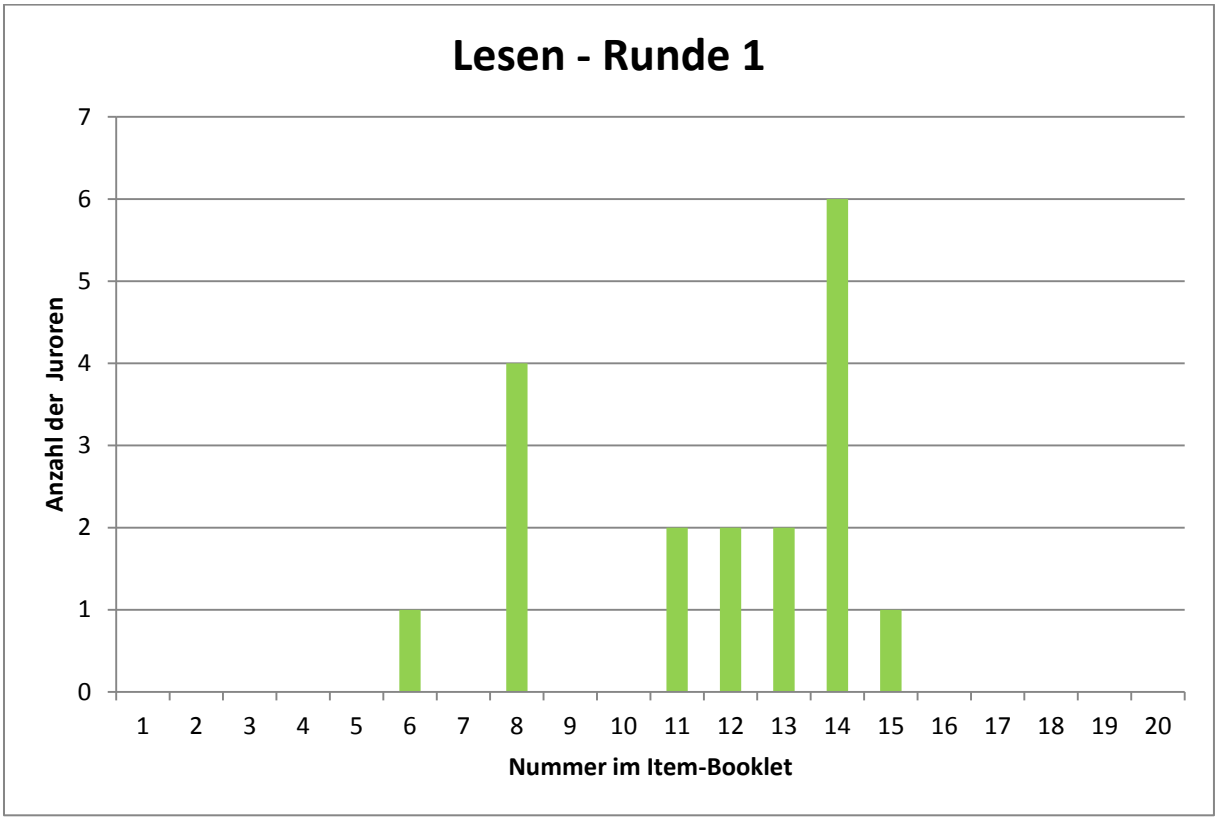


Schaubild 1 zeigt, welches Item als dasjenige in Runde 1 markiert wurde, das Prüfungsteilnehmende gerade noch erfolgreich bewältigen müssen, um damit Kenntnisse auf Niveau A2 nachweisen zu können. Die Bandbreite der Markierungen ist in Runde 1 groß und reicht von Item 6 bis Item 15.

Schaubild 2

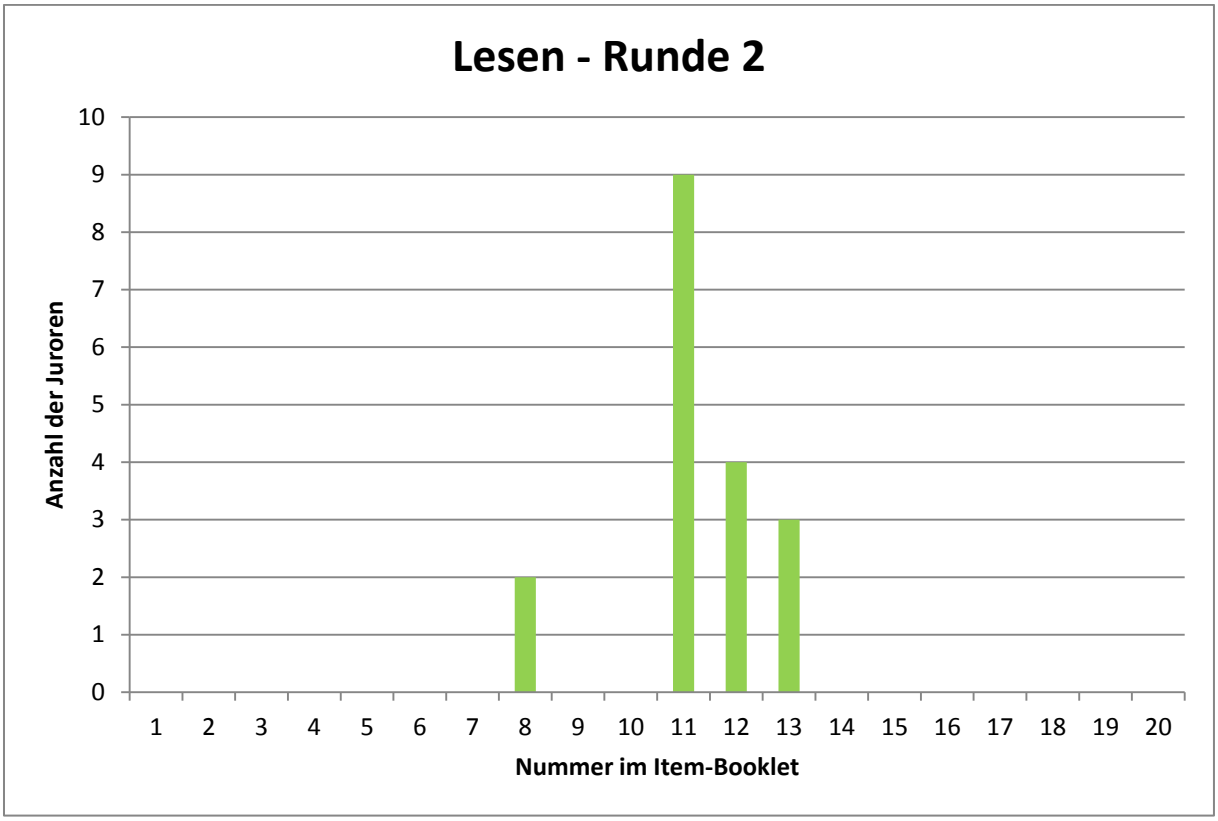


Schaubild 2 zeigt die Ergebnisse der Runde 2, die **nach** der Diskussion stattfand. Die Bandbreite der Markierungen hat sich deutlich verringert. Die Markierung für das Bestehen (cut score) für den Prüfungsteil *Lesen* wurde in Runde 1 bei dem Durchschnittswert (mean) 11,6 gesetzt, in Runde 2 bei dem Durchschnittswert 11,2. Es fand somit nach der Diskussion eine geringfügige Angleichung der Bewertungen statt.

2.2 Hören

Leitung: Ekaterini Karamichali

Assistenz: Claudia Stelter, Falko Röhrs

Teilnehmende: Vera Beiser-Kolb, Landesverband der VHS Saarland
Giulia Comparato, Klett-Langenscheidt Verlag
Stefanie Dengler, Goethe-Institut Zentrale, Bereich 41
Katharina Heydenreich, Spotlight Verlag
Marco Kellermann, Goethe-Institut Mannheim
Hildegard Kirchner, Goethe-Institut Freiburg
Jane Lloyd, Cambridge English Language Assessment
Sylke Loew, Sprachenzentrum der Universität des Saarlandes
Waldemar Martyniuk, Jagellionen-Universität in Krakau
Enikö Rabl, Ernst Klett Sprachen
Arthur Rapp, Goethe-Institut Zentrale, Bereich 42
Christiane Schmid, SDI München
Naomi Shafer, Universität Freiburg/Université de Fribourg
Kathrin Sokolowski, Cornelsen Schulverlage
Nora Tahy, Hueber Verlag
Brigitte Widmann, Freie Universität Bozen
Heike Widmer-Behr, Zürcher Hochschule für Angewandte Wissenschaften
Sonja Zimmermann, TestDaF-Institut

2.2.1 Verfahren

Wie auch in den anderen Arbeitsgruppen war es in dieser Gruppe das Ziel, die Bestehensgrenze des Prüfungsteils *Hören* zu bestimmen. Auch hier wurde dazu in einem mehrstufigen Verfahren ein Leistungsstandard als kritischer Wert (cut score) festgelegt, um sicherzustellen, dass die Anforderungen der Prüfung *Goethe-Zertifikat A2* (für Jugendliche und Erwachsene) im Prüfungsteil *Hören* dem angezielten Niveau A2 des *Gemeinsamen europäischen Referenzrahmens für Sprachen (GeR)* entsprechen.

Auch in dieser Arbeitsgruppe wurde der cut-off mit der Bookmark-Methode bestimmt. Nähere Erläuterungen zu dieser Methode und zum Vorgehen finden Sie unter 2.1.1.

Die Items im Booklet zum Prüfungsteil *Hören* stammen sowohl aus dem Modellsatz für Erwachsene (Items 5, 6, 8, 10, 12–15, 16, 18, 20) als auch aus dem für Jugendliche (1–4, 7, 9,

11, 17, 18). Die Modellsätze finden sich auf der Homepage des Goethe-Instituts (siehe Impressum).

2.2.2 Ergebnisse

Schaubild 1

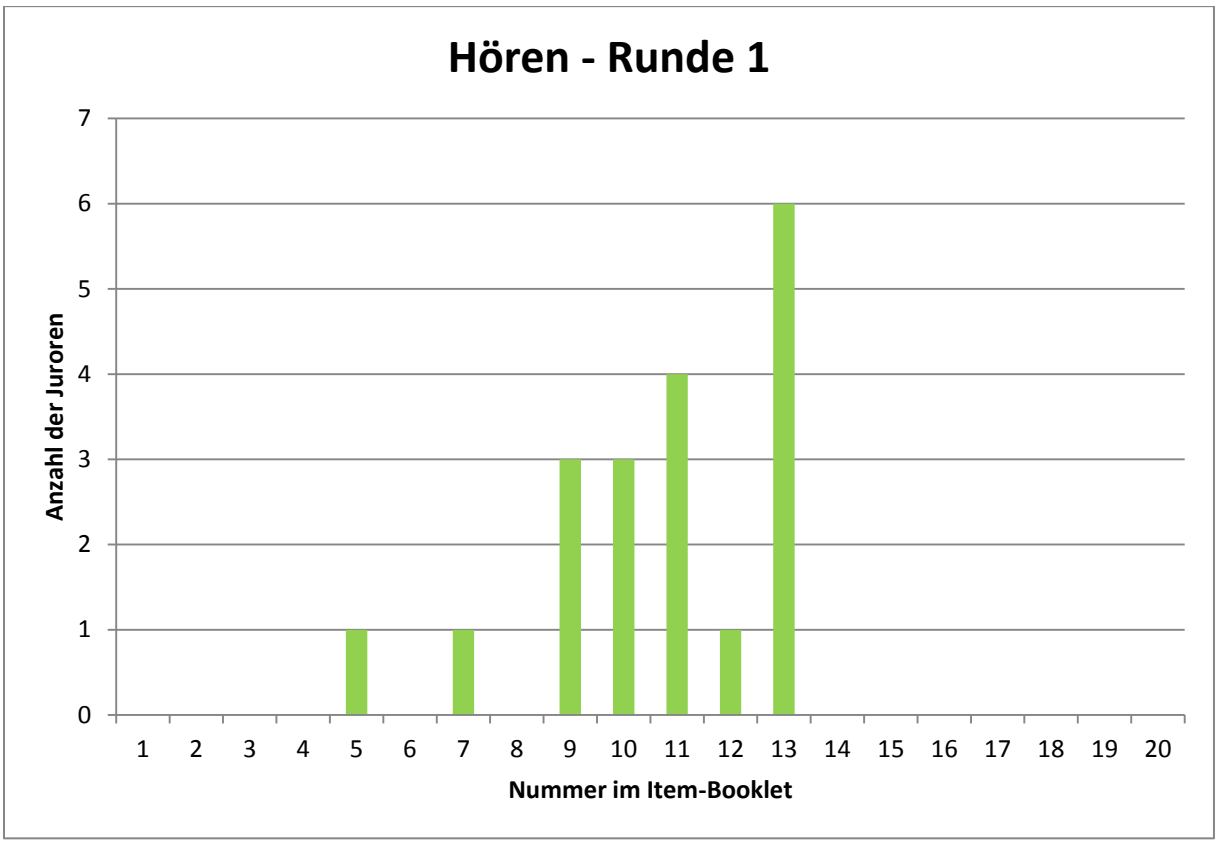


Schaubild 1 zeigt, welches Item die Jurorinnen und Juroren als dasjenige in Runde 1 und **vor** der Diskussion markiert haben, das Prüfungsteilnehmende erfolgreich bewältigen müssen, um damit Kenntnisse auf Niveau A2 nachweisen zu können. Die Bandbreite der Markierungen ist in Runde 1 relativ groß; sie reicht von Item 5 bis Item 13.

Schaubild 2

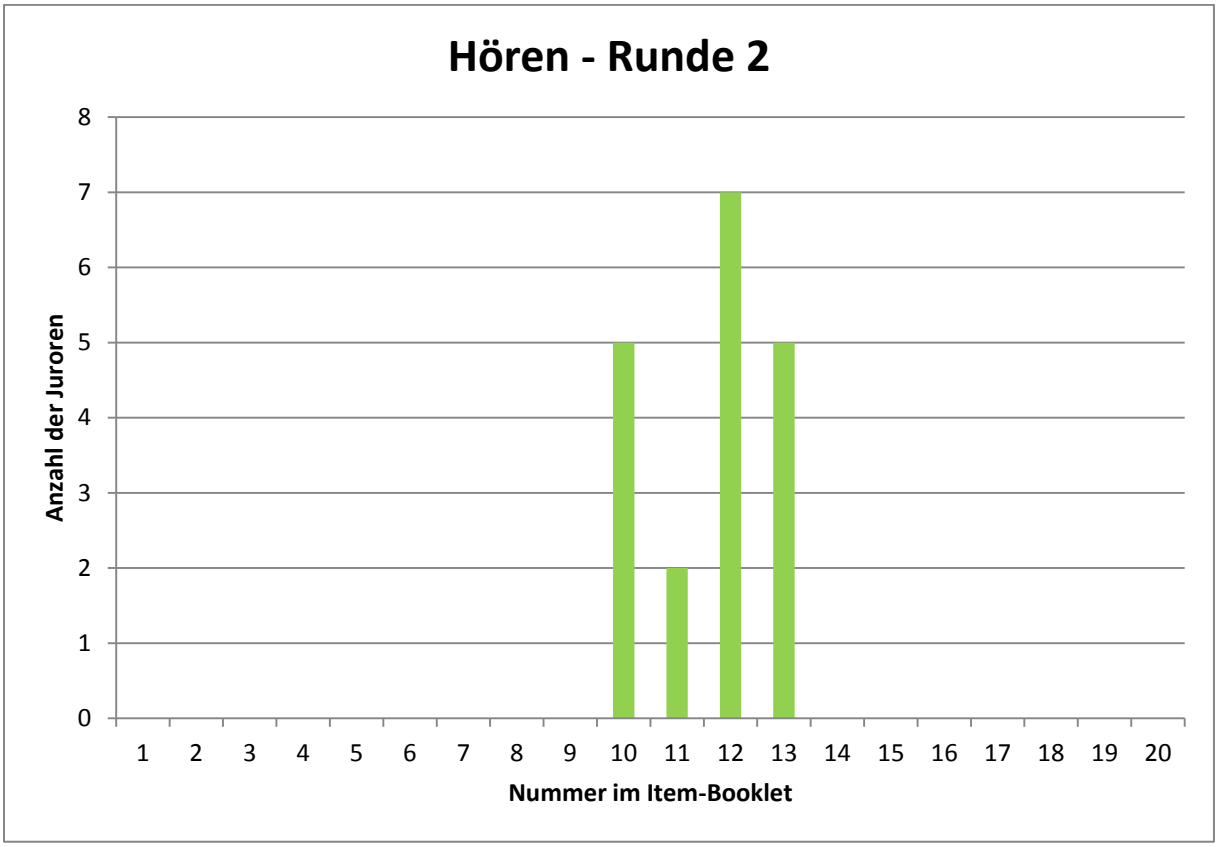


Schaubild 2 zeigt die Ergebnisse in Runde 2, **nach** der Diskussion der Jurorinnen und Juroren. Die Bandbreite der Markierungen hat sich deutlich verringert. Die Markierung für das Bestehen (cut score) für den Prüfungsteil *Hören* wurde in Runde 1 bei dem Durchschnittswert (mean) 10,6 gesetzt, in Runde 2 bei dem Durchschnittswert 11,6. Es fand also nach der Diskussion eine Angleichung der Bewertungen statt.

2.3 Schreiben

Leitung: Michaela Perlmann-Balme

Assistenz: Doris Hennemann, Linda Fromme

Teilnehmende: Ulrike Arras, Freie Universität Bozen
Gudula Bieber-Reynartz, Münchner Volkshochschule
Dominik Breithaupt, SDI München
Kirsten Bröcker, Landesverband der VHS Sachsen-Anhalt
Silvia Demmig, Friedrich-Schiller-Universität Jena
Armin Göbels, Goethe-Institut Berlin
Dorrie Goossens, Cito Niederlande
Corinna Hilger, Cornelsen Schulverlage
Ina Hoischen, Goethe-Institut Zentrale, Bereich 42
Silke Jacobs, Goethe-Institut Düsseldorf
Tanja Krüger, Goethe-Institut Zentrale, Bereich 41
Stefan Laub, did deutsch-institut
Uta Loumiotis, Klett Hellas
Florian Nimmrichter, Österreichisches Sprachdiplom
Stefanie Plisch de Vega, Ernst Klett Sprachen
Annerose Remus, Klett-Langenscheidt Verlag
Irmingard Staudigel, Landesverband der VHS Bayern
Virginia Suter Reich, Zürcher Hochschule für Angewandte Wissenschaften
Elisabetta Terrasi-Haufe, Ludwig-Maximilians-Universität München

2.3.1 Verfahren

Ziel dieser Arbeitsgruppe war es, nachzuweisen, dass die auf der Basis der Aufgaben erhobenen Teilnehmerleistungen im Prüfungsteil *Schreiben* mit der Definition des angestrebten Niveaus im *Gemeinsamen europäischen Referenzrahmen für Sprachen (GeR)* kompatibel sind. Ein weiteres Ziel bestand darin, eine Reihe von Referenzleistungen zu erhalten, die von Expertinnen und Experten auf dem Niveau A2 verortet wurden.

Zunächst wurden die teilnehmenden Jurorinnen und Juroren mithilfe der Deskriptoren des *Referenzrahmens* aus Kapitel 4 *Schriftliche Produktion Allgemein, Schriftliche Interaktion Allgemein* und aus Kapitel 5 *Spektrum sprachlicher Mittel allgemein, Kohärenz und Kohäsion*, *Wortschatzspektrum*, *Wortschatzbeherrschung, Grammatische Korrektheit* mit dem Niveau

A2 sowie den Nachbarniveaus vertraut gemacht. Ein solches Vertrautmachen war insofern notwendig, als sich die Einstufung allein auf diese Deskriptoren stützt und nicht etwa auf Bewertungskriterien zur Prüfung. Anschließend wurden drei Vergleichsarbeiten auf dem Niveau A1, A2 und B1 aus dem Material des Europarates (2005) herangezogen, um das Leistungsniveau A2 zu identifizieren und die Einstufung zu trainieren. Danach wurden zu jeder der beiden Aufgaben je 20 Leistungsbeispiele für Erwachsene und 20 für Jugendliche bearbeitet. Die Jurorinnen und Juroren entschieden, welche Leistungsbeispiele zum Schreiben auf der Niveaustufe A2 oder darüber zu verorten sind bzw. ob das Niveau A2 erreicht wurde. Sie gaben ihre Urteile anonym ab, d. h. jede bzw. jeder hatte eine Identifikationsnummer.

Die Schreibanlässe waren:

- Aufgabe 1: *Verabredung*
- Aufgabe 2: *Einladung*

Die Aufgaben finden sich auf der Homepage des Goethe-Instituts (siehe Impressum).

Die Arbeit vollzog sich in zwei Runden. In Runde 1 wurden zu den Aufgaben 1 und 2 jeweils 20 von erwachsenen und jugendlichen Erprobungsteilnehmenden verfasste Texte eingestuft. Dabei waren die Beispiele der Erwachsenen formal durch Maschinenschrift vereinheitlicht, die Jugendbeispiele blieben im Scan handschriftlich. Diese Leistungsbeispiele wurden zuerst in Einzelarbeit bearbeitet. Grundlage der Beurteilung waren die Deskriptoren des *Referenzrahmens*. Die Ergebnisse wurden aufgezeichnet, die Auswertung in der Gruppe präsentiert (vgl. Schaubild 1 bis 7). In der Übersicht war es den Jurorinnen und Juroren möglich, ihre Einstufungen über ihre Identifikationsnummer wiederzufinden und mit denen der anderen zu vergleichen. Drei Teilgruppen von jeweils sechs bzw. sieben Jurorinnen und Juroren wurden gebildet, wobei in jeder Gruppe strenge und milde vertreten waren. Ziel dieser Gruppendiskussion war es, die Einzelvoten zu begründen und diejenigen, deren Werte stärker vom Rest der Teilgruppe abwichen, zu einer Reflexion zu bringen. Ein Gruppenkonsens war nicht erforderlich. Nach Abschluss der Diskussion wurde für die Leistungsbeispiele der ersten Runde von jeder bzw. jedem einzeln ein zweites Votum abgegeben. In Runde 2 wurden je 20 weitere Leistungsbeispiele bewertet, jedoch aus Zeitgründen nur für die Aufgabe 1. Die Einstufung wurde nur in Einzelarbeit vorgenommen.

2.3.2 Ergebnisse

Die **Schaubilder 1 bis 6** zeigen die Globaleinstufung der Leistungen

0 = unterhalb Niveau A2

1 = Niveau A2 oder darüber

Sie zeigen, wie viele Personen ein Leistungsbeispiel als auf A2 liegend bewertet haben. Auf der horizontalen Achse befinden sich oben die Beispiele 1 bis 20, unten die erzielten Ergebnisse pro Beispiel. Auf der vertikalen Achse sind links die 19 Jurorinnen und Juroren (Rater) aufgelistet, rechts die Gesamtzahl der von den Jurorinnen und Juroren auf A2 eingestuften Beispiele. Die Schaubilder 1 und 2 zeigen die Ergebnisse **vor** der Diskussion.

Schaubild 1

Verabredung (Erwachsene)																						
		Beispiel																				Ergebnis Rater %
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Rater	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	0	0	1	1	1	0	50%
	2	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	85%
	3	1	1	0	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	80%
	4	1	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	60%
	5	1	0	0	1	1	1	1	1	0	1	0	0	0	1	0	0	1	1	1	0	55%
	7	1	0	0	1	1	1	1	0	0	1	0	0	0	0	0	0	1	1	0	1	45%
	8	1	0	0	1	1	0	1	0	1	1	0	0	1	1	0	0	1	1	1	1	60%
	9	1	0	0	1	1	0	1	0	0	1	0	0	0	0	0	0	1	1	1	0	40%
	10	1	0	0	1	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	1	65%
	11	1	1	0	1	1	0	1	0	1	1	1	0	1	1	0	0	1	1	1	1	70%
	12	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	80%
	13	1	0	0	1	1	0	1	0	1	1	0	0	1	0	0	0	1	1	0	1	50%
	14	1	0	0	1	1	0	1	0	1	1	0	0	0	0	0	0	1	1	0	0	40%
	15	1	0	0	1	1	1	1	1	1	1	1	0	1	0	0	0	1	1	1	1	70%
	16	1	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	70%
	17	1	0	0	1	1	0	1	0	1	1	0	0	1	0	0	0	1	1	0	1	50%
	18	1	0	0	1	1	0	1	0	1	1	1	0	1	0	0	0	1	1	1	0	55%
	19	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	1	1	1	75%
	20	1	0	0	1	1	0	1	0	1	1	0	1	1	0	0	0	1	1	1	1	60%
Ergebnis Aufgabe %		100%	32%	11%	100%	100%	53%	100%	47%	84%	95%	32%	26%	63%	26%	0%	0%	100%	100%	79%	74%	

Schaubild 1: Die Beispiele 1, 4, 5, 7, 17 und 18 wurden übereinstimmend auf Niveau A2 oder darüber eingestuft, die Beispiele 15 und 16 wurden von allen unter Niveau A2 bewertet. Bei den restlichen Beispielen gab es kein eindeutiges Votum. Dabei waren die Beispiele 6 und 8 besonders strittig.

Hinweise: Die Nummern der Rater entsprechen nicht der Abfolge der Teilnehmenden unter 2.3.1. In den Schaubildern 1 bis 10 in diesem Kapitel fehlt jeweils der/die Juror/-in mit der Nummer 6, da er/sie kurzfristig verhindert war, an der Konferenz teilzunehmen.

Schaubild 2

Verabredung (Jugendliche)																					
		Beispiel																			Ergebnis Rater %
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Rater	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	0	80%
	2	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	80%
	3	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	90%
	4	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	0	75%
	5	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	80%
	7	0	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	80%
	8	1	1	0	0	1	0	1	1	1	1	1	0	1	1	0	1	1	1	0	1	70%
	9	1	1	0	1	1	0	1	0	1	1	1	0	1	1	1	1	1	0	1	0	70%
	10	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	85%
	11	1	1	0	1	1	0	1	1	1	1	1	0	1	0	1	1	1	0	1	1	75%
	12	1	1	0	1	1	0	1	1	1	1	1	0	1	0	1	1	1	0	1	1	75%
	13	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	0	1	0	1	1	80%
	14	1	1	0	1	1	0	1	1	1	1	1	0	1	1	0	1	1	0	1	0	70%
	15	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	90%
	16	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	90%
	17	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	0	75%
	18	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	0	80%
	19	1	1	0	1	1	0	1	1	1	1	1	0	1	1	0	1	1	0	1	0	70%
	20	1	1	0	1	1	0	0	1	1	1	1	0	1	1	0	1	1	0	1	0	65%
Ergebnis Aufgabe %		95%	100%	0%	95%	100%	32%	95%	95%	100%	100%	100%	11%	100%	89%	74%	95%	100%	42%	95%	42%	

Schaubild 2 Die Anzahl von 100-prozentigen Übereinstimmungen (100 bzw. 0 %) lag hier ebenfalls bei acht. Allerdings war die Gruppe der weitgehenden Übereinstimmungen (95 % und 11 %) mit sieben höher.

Schaubild 3

Einladung (Erwachsene)																					
		Beispiel																			Ergebnis Rater %
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Rater	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	0	85%
	2	1	1	1	0	0	0	0	1	1	1	1	0	1	0	1	1	1	1	1	0	65%
	3	1	1	1	1	1	0	0	1	1	0	1	0	1	0	1	1	1	1	1	0	70%
	4	1	1	0	0	1	0	0	1	0	0	0	0	1	0	1	1	1	1	1	0	50%
	5	1	1	1	1	1	0	0	1	1	0	1	0	0	0	1	1	1	1	1	0	65%
	7	1	1	0	0	0	1	0	1	1	0	1	0	1	0	1	1	1	1	1	0	60%
	8	1	1	1	1	1	0	0	1	0	1	1	0	1	0	1	1	1	1	1	0	70%
	9	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	30%
	10	1	1	1	0	0	1	0	1	1	0	1	0	1	0	1	1	1	1	1	0	65%
	11	1	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	1	0	75%
	12	1	0	1	0	0	1	0	0	1	0	0	0	1	1	1	1	1	1	1	0	55%
	13	1	1	1	0	0	0	0	0	0	1	1	0	1	0	1	1	1	1	1	0	55%
	14	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	0	45%
	15	1	1	1	1	0	0	0	1	0	0	0	0	0	0	1	1	1	1	1	0	50%
	16	1	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	1	1	1	0	40%
	17	1	1	1	0	0	0	0	1	1	1	1	0	0	0	1	1	1	1	1	0	60%
	18	1	1	1	1	0	0	0	1	0	0	1	0	1	0	1	1	1	1	1	0	60%
	19	1	1	1	1	0	1	0	1	0	0	1	1	1	0	1	1	1	1	1	0	70%
	20	1	1	1	1	0	1	0	1	1	1	1	0	1	0	1	1	1	1	1	0	75%
Ergebnis Aufgabe %		100%	95%	84%	47%	32%	37%	0%	79%	58%	32%	63%	5%	68%	16%	95%	95%	100%	100%	100%	0%	

Schaubild 4

Einladung (Jugendliche)																						
		Beispiel																				Ergebnis Rater %
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Rater	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	85%
	2	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	1	0	1	1	0	70%
	3	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	90%
	4	1	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1	0	1	1	0	75%
	5	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	0	1	1	0	75%
	7	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	1	0	80%
	8	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	1	0	80%
	9	1	1	1	0	0	1	1	1	1	0	1	1	0	0	1	1	0	1	1	0	65%
	10	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	1	1	1	80%
	11	1	1	1	0	1	1	1	1	1	0	1	1	0	0	1	1	0	1	1	0	70%
	12	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	1	1	0	75%
	13	1	1	0	1	1	1	1	1	1	0	1	1	0	0	1	1	0	0	1	0	65%
	14	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	1	0	1	1	0	70%
	15	1	1	0	0	0	1	1	1	1	0	1	1	0	0	1	1	0	1	1	0	60%
	16	1	1	0	0	0	1	1	1	1	0	1	1	0	0	1	0	0	0	1	0	50%
	17	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	1	0	1	1	0	70%
	18	1	1	1	0	1	1	1	1	1	0	1	1	0	0	1	1	0	1	1	0	70%
	19	1	1	1	0	0	1	1	1	1	0	1	1	1	0	1	1	0	1	1	0	70%
	20	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1	0	80%
Ergebnis Aufgabe %		100%	100%	79%	68%	58%	95%	100%	100%	100%	21%	100%	100%	26%	0%	100%	95%	16%	89%	100%	5%	

Schaubild 3 und 4 zeigen die Ergebnisse für Aufgabe 2 *Einladung* (1. Runde) Erwachsene und Jugendliche **vor** der Diskussion. Bei den Erwachsenen wurde zu sechs Beispielen eine totale, in weiteren vier eine weitreichende Übereinstimmung erzielt.

Schaubild 5

Verabredung (Erwachsene)																						
		Beispiel																				Ergebnis Rater %
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Rater	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	0	55%
	2	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	1	1	1	1	75%
	3	1	1	0	1	1	0	1	1	1	1	0	1	0	1	0	0	1	1	1	1	70%
	4	1	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	60%
	5	1	0	0	1	1	1	1	1	0	1	0	0	0	1	0	0	1	1	1	0	55%
	7	1	0	0	1	1	0	1	0	0	1	1	0	1	0	0	0	1	1	0	1	50%
	8	1	0	0	1	1	0	1	0	1	1	0	0	0	1	0	0	1	1	1	1	55%
	9	1	0	0	1	1	1	1	1	0	1	0	0	0	0	0	0	1	1	1	0	50%
	10	1	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	60%
	11	1	1	0	1	1	0	1	0	1	1	0	0	1	1	0	0	1	1	1	1	65%
	12	1	1	0	1	1	0	1	0	1	1	0	0	1	1	0	0	1	1	1	1	65%
	13	1	0	0	1	1	0	1	1	1	1	0	0	0	0	0	0	1	1	0	1	50%
	14	1	0	0	1	1	0	1	1	1	1	0	0	0	0	0	0	1	1	0	0	45%
	15	1	0	0	1	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	1	65%
	16	1	0	0	1	1	1	1	1	1	1	0	1	0	0	0	0	1	1	1	1	65%
	17	1	0	0	1	1	0	1	0	1	1	0	0	1	0	0	0	1	1	0	1	50%
	18	1	0	0	1	1	0	1	0	1	1	0	0	1	0	0	0	1	1	1	0	50%
	19	1	0	1	1	1	0	1	0	1	1	0	1	1	0	0	0	1	1	1	1	65%
	20	1	0	0	1	1	0	1	0	1	1	0	1	0	0	0	0	1	1	1	1	55%
Ergebnis Aufgabe %		100%	26%	11%	100%	100%	42%	100%	58%	84%	95%	5%	26%	37%	26%	0%	0%	100%	100%	79%	74%	

Schaubild 6

Verabredung (Jugendliche)																					
		Beispiel																			Ergebnis Rater %
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Rater	1	1	1	0	1	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	0	75%
	2	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	80%
	3	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	90%
	4	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	0	75%
	5	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	80%
	7	0	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	75%
	8	1	1	0	0	1	0	1	1	1	1	1	0	1	1	0	1	1	1	0	1	70%
	9	1	1	0	1	1	0	1	0	1	1	1	0	1	1	1	1	1	0	1	0	70%
	10	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	85%
	11	1	1	0	1	1	0	1	1	1	1	1	0	1	0	1	1	1	0	1	1	75%
	12	1	1	0	1	1	0	1	1	1	1	1	0	1	0	1	1	1	0	1	1	75%
	13	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1		1	0	1	1	80%
	14	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	80%
	15	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	85%
	16	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	85%
	17	1	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	80%
	18	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	85%
	19	1	1	0	1	1	0	1	1	1	1	1	0	1	1	0	1	1	0	1	1	75%
	20	1	1	0	1	1	0	0	1	1	1	1	0	1	1	0	1	1	0	1	0	65%
Ergebnis Aufgabe %		95%	100%	0%	95%	100%	16%	95%	95%	100%	100%	100%	11%	100%	89%	79%	95%	100%	37%	95%	63%	

Schaubild 5 und 6 zeigt die Ergebnisse für Aufgabe 1 *Verabredung* (2. Runde) **nach** der Diskussion. Die Zahl der 100-prozentigen Übereinstimmungen wuchs durch die Diskussion in beiden Fällen nicht an.

2.3.3 Statistische Analysen

Die statistischen Analysen mit dem Programm FACETS erbrachten Erkenntnisse über die Beispiele (sample), Jurorinnen und Juroren (rater) und Aufgaben (task). Die Ergebnisse wurden mit der Maßeinheit logits dargestellt. Die Beispiele zu Aufgabe 1 *Verabredung* tragen die Kennzeichnung V oder E für *Einladung*. Der Kleinbuchstabe e bzw. j steht für Erwachsene oder Jugendliche. Die Ziffer davor bezeichnet die Nummer des Beispiels.

Schaubild 7

Beispiele, Rater, Aufgaben

In Runde 1 erreichten die Beispiele mit den höchsten Werten +4 logits. Diese Beispiele wurden von den meisten auf A2 eingestuft Die Beispiele mit den niedrigsten Werten erreichten −4 logits. Diese Beispiele wurden von den meisten als unterhalb A2 eingestuft. Die Rater 9 und 14 waren in ihrer Einstufung strenger als die anderen, Rater 3 war der mildeste. Die Mehrheit findet sich bei +0,5 und −0,5 logits und ist damit weder streng noch mild. Die beiden Aufgaben wurden als etwa gleich schwer eingestuft, die *Einladung* etwas schwerer als die *Verabredung*.

Measr	Sample	rater	status	task
4	10Vj 11Ej 11Vj 12Ej 13Vj 15Ej 17Ee 17Ve 17Vj 18Ee 18Ve 19Ee 19Ej 1Ee 1Ej 1Ve 2Ej 2Vj 4Ve 5Ve 5Vj 7Ej 7Ve 8Ej 9Ej 9Vj			
	15Ee 16Ee 16Ej 2Ee 6Ej			
3	10Ve 16Vj 19Vj 1Vj 4Vj 7Vj 8Vj			
	18Ej			
	14Vj			
2	3Ee	r9		
	3Ej 8Ee 9Ve			
	19Ve	r14		
1	13Ee 15Vj 20Ve 4Ej			
	11Ee			
		r13 r16		
	13Ve 5Ej 9Ee	r17		
		r4		
		r18 r7		Einladung
0	4Ee 6Ve	r15 r5	Runde 1	
		r20 r8		Verabredung
	8Ve	r12 r19		
	18Vj 20Vj 6Ee	r11		
	10Ee 5Ee	r10		
		r1 r2		
-1	11Ve 13Ej 2Ve 6Vj			
	10Ej 12Ve 14Ve			
	14Ee 17Ej			
-2		r3		
	12Vj 3Ve			
-3	12Ee 20Ej			
-4	14Ej 15Ve 16Ve 20Ee 3Vj 7Ee			

Schaubild 8

Für Runde 2 wurde dieselbe Analyse für die Aufgabe Verabredung vorgenommen. Die Bandbreite bei der Verteilung der Beispiele zwischen +4 und −4 logits hat sich nicht verändert. Das bedeutet, dass die Diskussion zwischen den beiden Runden zu keinen Veränderungen bezüglich der Einstufung der Beispiele führte. Ein Vergleich der beiden Schaubilder erlaubt es, die Bewegungen einzelner Jurorinnen und Juroren bezüglich Milde bzw. Strenge ihrer Bewertungen nachzuverfolgen. Rater 3 ist weiterhin der mildeste und Rater 9 weiterhin einer der strengsten. Abgesehen von Rater 3 und 2 fallen alle jetzt in die Gruppe zwischen +1 und −1 logits. Die Strenge der Jurorinnen und Juroren reicht in der Runde 1 von 2 logits für Rater 9 bis −2 logits für Rater 3, in Runde 2 rückten diese Werte näher zusammen auf 1 logit für Rater 9 und −1,6 logits für Rater 3.

Measr	Sample	rater	status	task
4	10Vj 11Vj 13Vj 17Ve 17Vj 18Ve 1Ve 2Vj 4Ve 5Ve 5Vj 7Ve 9Vj			
	10Ve 16Vj 19Vj 1Vj 4Vj 7Vj 8Vj			
3				
	14Vj			
2				
	9Ve			
	15Vj 19Ve			
	20Ve			
1		r20 r9		
	20Vj	r14 r7		
		r1 r13		
	8Ve			
		r18 r4		
0			Runde 2	Verabredung
		r11 r12		
	6Ve			
		r10		
	13Ve 18Vj			
		r15 r16		
-1				
	12Ve 14Ve 2Ve			
		r2		
		r3		
	6Vj			
-2				
	12Vj 3Ve			
-3				
	11Ve			
-4	15Ve 16Ve 3Vj			

Schaubild 9

Intra-Rater-Zuverlässigkeit

Da die Ergebnisse dieses Benchmarkings urteilsbasierter Natur sind ist) eine Evaluation der Zuverlässigkeit der Jurorinnen und Juroren vorzunehmen. Analysiert wurde dazu die Differenz zwischen den in Runde 1 und Runde 2 abgegebenen Werten (0 oder 1). Mithilfe von Alpha (Spalte 2) wurde die Zuverlässigkeit bestimmt. Jemand, der in beiden Runden dieselbe Anzahl an 0- und 1-Werten vergab, erhielt den Wert 1.000 für komplette Übereinstimmung. Denselben Wert erhielt, wer zweimal identische Werte abgab. Kappa (Spalte 3) misst exakte Übereinstimmung und zeigt, in welchem Maße die Werte von Runde 1 und 2 korrelieren. Ein/-e Juror/-in, der/die in beiden Runden dieselben Werte vergab (z. B. r4 und r5) erhielt einen Kappa-Wert 1.000. Der Wert in Spalte 4, Pearson, identifiziert Rater, die konsistent Werte beibehalten. Sie haben einen höheren Wert als Rater, die von Runde zu Runde mal streng und mal mild bewerten. Drei von 20 Jurorinnen und Juroren im Schaubild zeigen stellenweise auffällige Werte. Nur Rater 17 zeigt in Bezug auf Zuverlässigkeit, Übereinstimmung und Konsistenz auffällige Werte. Dieses Gesamtergebnis untermauert die Glaubwürdigkeit im Benchmarking erzielten Ergebnisse.

rater	Alpha/ICC	Kappa	Pearson
r1	0.942	0.647	0.722
r2	0.919	0.844	0.855
r3	0.910	0.828	0.840
r4	1.000	1.000	1.000
r5	1.000	1.000	1.000
r7	0.881	0.787	0.787
r8	0.973	0.946	0.947
r9	0.949	0.898	0.903
r10	0.967	0.935	0.937
r11	0.969	0.939	0.941
r12	0.901	0.808	0.823
r13	0.942	0.890	0.890
r14	0.922	0.846	0.856
r15	0.927	0.857	0.866
r16	0.927	0.857	0.866
r17	0.753	0.595	0.605
r18	0.940	0.886	0.886
r19	0.900	0.817	0.819
r20	0.974	0.947	0.949

Schaubild 10

Inter-Rater-Zuverlässigkeit

Verglichen wurden die erwartete Übereinstimmung zwischen Jurorinnen und Juroren mit den tatsächlich beobachteten Übereinstimmungen, bezogen auf 13.680 mögliche Übereinstimmungen in Runde 1 und 6.840 in Runde 2.

	erwartet	**beobachtet**
Runde 1	81.1%	81.9%
Runde 2	82.9%	83.7%

Die vorliegende Nähe zwischen erwarteter und beobachteter Übereinstimmung ist ein Indikator für eine sehr gute Inter-Rater-Zuverlässigkeit. Im vorliegenden Daten-Set waren 2 % der Beobachtungen in der Runde 1 unerwartet, in der 2. Runde waren 3 % der Beobachtungen unerwartet. Die beteiligten Juror/-innen waren daher sehr zuverlässig.

2.3.4 Bewertete Leistungsbeispiele

Leistungsbeispiele, die klar **auf** Niveau A2 angesiedelt wurden.

Beispiel 1Ve	**Beispiel 1Ej**
Aufgabe 1: Verabredung Erwachsene	Aufgabe 2: Einladung Jugendliche
Hallo Ekaterini, Tut mir leid! Ich bin im Stau. Komme später. Treffen uns um 7 Uhr auf Hauptbahnhof. Bis später.	Entschuldigung Ekaterini! Ich war in den Versammlung, deshalb komme ich spät. Können wir am sechs Uhr vor der Bank treffen. Es tut mir Leid noch ein mal.

Leistungsbeispiele, die klar **unterhalb** Niveau A2 angesiedelt wurden.

Beispiel 15Ve	**Beispiel 20Ej**
Aufgabe 1: Verabredung Erwachsene	Aufgabe 2: Einladung Jugendliche
Der Zug ist Kapput. hate nicht viel Zeit. Hate kein Information. Gehen sie zu Teather, oder Ich hoffe zugehe schnell.	im Freitag. Ich habe fußballmatch Freitag en schuldigun!!! Das fußallmatch sehr sehimportant . Das ist fina

Leistungsbeispiele, die mit circa 50-prozentiger Wahrscheinlichkeit auf Niveau A2 angesiedelt wurden.

Beispiel 6Ve	**Beispiel 5Ej**
Aufgabe 1: Verabredung Erwachsene	Aufgabe 2: Einladung Jugendliche
Liebe Ekaterini, Ich komme Halb Uhr später. Ich vergiss mein documents zu Hause. Sagt du, können wir um 9.40 Uhr im Kafe treffen?	Guten Tag! Vielen Dank! Ich komme. Eine Party? Das finde ich super! Aber ich weiss nicht wo das Sportzentrum ist. Können Sie zu mich antworten? Bitte bald! Danke. Viele Grüße, Deine Vanessa.

Schaubild 11

Vergleich Benchmarking und Trainingsmaterialien

	Aufgabe 1 Erwachsene	
Beispiel	Benchmarking	Training
1	A2 100 %	Nr. 1 A2
10	A2 95 %	Nr. 10 A2
15	Unter A2 0 %	Nr. 5 Unter A2

	Aufgabe 2 Erwachsene	
Beispiel	Benchmarking	Training
2	A2 95 %	Nr. 6 A2
4	Unter A2 47 %	Nr. 7 Unter A2
10	A2 32 %	Nr. 11 Unter A2

	Aufgabe 1 Jugendliche	
Beispiel	Benchmarking	Training
16	A2 95 %	Nr. 3
18	Unter A2 37 %	Nr. 4
1	A2 95 %	Nr. 9

	Aufgabe 2 Jugendliche	
Beispiel	Benchmarking	Training
1	A2 100 %	Nr. 2
20	Unter A2 5%	Nr. 8
13	Unter A2 26 %	Nr. 12

Schaubild 11 geht auf die Verbindung zwischen der Niveaueinstufung des Benchmarkings und den Ergebnissen der Bewertung durch Anwendung der für die Prüfung entwickelten Bewertungskriterien ein. Diese Ergebnisse finden sich in dem *Trainingsmaterial für Prüfende*. Die Gegenüberstellung zeigt, dass die Juror/-innen des Benchmarkings zu denselben Schlüssen gelangten wie die Bewertenden, die die Leistungsbeispiele im Trainingsmaterial bewerteten.

2.4 Sprechen

Leitung: Claudia Stelter

Assistenz: Ekaterini Karamichali, Falko Röhrs

Teilnehmende: Vera Beiser-Kolb, Landesverband der VHS Saarland
Giulia Comparato, Klett-Langenscheidt Verlag
Stefanie Dengler, Goethe-Institut Zentrale, Bereich 41
Katharina Heydenreich, Spotlight Verlag
Marco Kellermann, Goethe-Institut Mannheim
Hildegard Kirchner, Goethe-Institut Freiburg
Jane Lloyd, Cambridge English Language Assessment
Sylke Loew, Sprachenzentrum der Universität des Saarlandes
Waldemar Martyniuk, Jagellionen-Universität in Krakau
Enikö Rabl, Ernst Klett Sprachen
Arthur Rapp, Goethe-Institut Zentrale, Bereich 42
Christiane Schmid, SDI München
Naomi Shafer, Universität Freiburg/Université de Fribourg
Kathrin Sokolowski, Cornelsen Schulverlage
Nora Tahy, Hueber Verlag
Brigitte Widmann, Freie Universität Bozen
Heike Widmer-Behr, Zürcher Hochschule für Angewandte Wissenschaften
Sonja Zimmermann, TestDaF-Institut

2.4.1 Ziel, Material

Ziel dieser Arbeitsgruppe war es zu überprüfen, ob die Aufgaben und die gefilmten Leistungsbeispiele zum Prüfungsteil *Sprechen* der Definition des A2-Niveaus im *Gemeinsamen europäischen Referenzrahmen für Sprachen (GeR)* entsprechen. Methodische Grundlage hierfür war das im Handbuch *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (2009) beschriebene dreistufige Vorgehen:

- Vertrautmachen mit den Deskriptoren, hier: „Mündliche Produktion allgemein“ (GeR S. 64), „Zusammenhängendes monologisches Sprechen: Erfahrungen beschreiben“ (GeR S. 64f) und „Mündliche Interaktion allgemein“ (GeR S. 79) sowie mit der Tabelle „Qualitative Aspekte des mündlichen Sprachgebrauchs“ (GeR S. 37f).

- Vorgabe je eines kalibrierten Beispiels zur Produktion und zur Interaktion, die im Auftrag des Europarats von Bewertenden eingestuft worden waren (Bolton et al., 2008; CIEP, 2008).
- Einstufung der Leistungsbeispiele zum Goethe-Zertifikat A2 nach dem Vertrautmachen mit der Aufgabenstellung.

Die gezeigten sechs Leistungsbeispiele waren 2014 und 2015 in der Zentrale des Goethe-Instituts und im Goethe-Institut Freiburg aufgenommen worden. Die Teilnehmenden lernten zu dieser Zeit Deutsch am Goethe-Institut München, der Volkshochschule München sowie im Jugendkurs am Goethe-Institut Freiburg.

Die Sprechanlässe in der Interaktion (Aufgabe 1 und 3) und Produktion (Aufgabe 2) waren:

- Aufgabe 1: Mithilfe von Stichworten Informationen zur Person mit einem Partner/einer Partnerin austauschen. *(Stichworte Erwachsene: Land, Hobby, Kinder, Sprachen, Beruf, Geburtstag, Freunde, Wohnort. Stichworte Jugendliche: Land, Hobby, Geschwister, Sprachen, Lieblingsfach, Geburtstag, Freunde, Wohnort)*
- Aufgabe 2: Dem/Der Prüfenden ausführlich nähere Informationen zum eigenen Leben geben. *(Erwachsene: Was machen Sie oft am Wochenende?* und *Was machen Sie mit Ihrem Geld? Jugendliche: Was machst du oft am Wochenende?* und *Was machst du mit deinem Taschengeld?)*
- Aufgabe 3: Mit einem Partner/einer Partnerin eine Unternehmung planen und aushandeln. *(Erwachsene: Ihr Freund Patrick hat Geburtstag. Sie möchten ein Geschenk für ihn kaufen. Finden Sie einen Termin. Jugendliche: Ihr wollt zusammen für Julias Geburtstagsfeier ein Geschenk kaufen. Findet einen Termin.)*

Diese Aufgaben finden sich auf der Homepage des Goethe-Instituts (siehe Impressum).

2.4.2 Verfahren

Die Jurorinnen und Juroren entschieden, welche der Leistungsbeispiele auf der Niveaustufe A2 oder darüber zu verorten sind bzw. welche das Niveau A2 nicht erreichten („unterhalb Niveau A2“). Die Urteile wurden anonym durch das Zuteilen einer Identifikationsnummer abgegeben. Die Juror/-innen entschieden zunächst jede/-r für sich. Ein Gruppenkonsens war nicht erforderlich. Die Ergebnisse der Niveaueinstufungen wurden dann bekannt gegeben. In der Übersicht war es den Jurorinnen und Juroren möglich, ihre Einstufungen über ihre Identifikationsnummer wiederzufinden und mit denen der anderen zu vergleichen. Es

erfolgte eine Diskussion in Kleingruppen über die Abweichungen. Die Kleingruppen wurden wie beim Prüfungsteil Schreiben heterogen aus milden und strengen Juror/-innen zusammengesetzt. Nach der Diskussion wurden die Leistungsbeispiele erneut angesehen und eingestuft.

2.4.3 Ergebnisse

Die Schaubilder 1 bis 6 zeigen die Globaleinstufung der Leistungen auf Basis der im *Gemeinsamen europäischen Referenzrahmen für Sprachen* festgelegten Deskriptoren wie folgt:

0 = unterhalb Niveau A2

1 = Niveau A2 oder darüber

Schaubild 1

Aufgabe 1- Interaktion		Beispiel						Ergebnis Rater %
		1e	2e	3e	4e	5j	6j	
Rater	1	1	1	1	1	1	1	100%
	2	1	1	1	1	1	1	100%
	3	1	1	1	1	1	1	100%
	4	1	1	1	1	1	1	100%
	6	1	1	1	1	1	1	100%
	7	1	1	1	1	1	1	100%
	8	1	1	1	0	1	1	83%
	9	1	0	1	0	1	1	67%
	10	1	1	1	0	1	1	83%
	11	1	1	1	0	1	1	83%
	12	1	1	1	1	1	1	100%
	13	1	1	1	1	1	1	100%
	14	1	1	1	0	1	1	83%
	15	1	1	1	1	1	1	100%
	16	1	1	1	1	1	1	100%
	17	1	1	1	1	1	1	100%
	19	1	1	1	0	1	1	83%
	20	1	1	1	1	1	1	100%
Ergebnis Aufgabe %		100%	94%	100%	67%	100%	100%	

Schaubild 1 zeigt die Ergebnisse zu Aufgabe 1 – Interaktion nach der ersten Bewertungsrunde, **vor** der Diskussion. Auf der horizontalen Achse befinden sich oben die Leistungsbeispiele 1 bis 6, gekennzeichnet mit *e* für ein Beispiel der Erwachsenenversion und mit *j* für ein Beispiel der Jugendversion. In der untersten Zeile finden sich die erzielten Ergebnisse pro Beispiel. 100 % bedeutet, dass alle 18 Jurorinnen und Juroren (Rater) das Beispiel für eine A2-Leistung hielten. Auf der vertikalen Achse sind links 18 Jurorinnen und Juroren aufgelistet. Die Nummern entsprechen nicht der Abfolge der Teilnehmende unter 2.4.1. In der Auflistung fehlen die Juror/-innen 5 und 18. Sie waren kurzfristig verhindert und konnten daher kein Votum abgeben. In der rechten Spalte sieht man, wie viele Beispiele insgesamt pro Juror/-in auf A2 eingestuft wurden. Die Beispiele 1, 3, 5, und 6 wurden von allen übereinstimmend auf Niveau A2 eingestuft. Bei Beispiel 2 gab es eine/-n, die/der diese Leistung unter dem Niveau A2 einstufte. Diskussionsbedarf gab es bei Beispiel 4, bei dem sechs für „unterhalb Niveau A2“ stimmten.

Schaubild 2

Aufgabe 1 - Interaktion		Beispiel						Ergebnis Rater %
		1e	2e	3e	4e	5j	6j	
Rater	1	1	1	1	1	1	1	100%
	2	1	1	1	1	1	1	100%
	3	1	1	1	1	1	1	100%
	4	1	1	1	1	1	1	100%
	6	1	1	1	1	1	1	100%
	7	1	1	1	1	1	1	100%
	8	1	1	1	0	1	1	83%
	9	1	1	1	1	1	1	100%
	10	1	1	1	1	1	1	100%
	11	1	1	1	1	1	1	100%
	12	1	1	1	0	1	1	83%
	13	1	1	1	1	1	1	100%
	14	1	1	1	1	1	1	100%
	15	1	1	1	1	1	1	100%
	16	1	1	1	1	1	1	100%
	17	1	1	1	1	1	1	100%
	19	1	1	1	0	1	1	83%
	20	1	1	1	1	1	1	100%
Ergebnis Aufgabe %		100%	100%	100%	83%	100%	100%	

Schaubild 2 zeigt die Ergebnisse zu Aufgabe 1 – Interaktion **nach** der Diskussion der Abweichungen und der zweiten Bewertungsrunde. Die Diskussion in Teilgruppen führte bei den Beispielen 2 und 4 zu einer stärkeren Einheitlichkeit des Votums. In fünf von sechs Beispielen wurde eine Übereinstimmung erzielt. Es wurde darauf verzichtet, die verbleibenden Abweichungen weiter zu diskutieren.

Schaubild 3

Aufgabe 2 - Produktion		Beispiel						Ergebnis Rater %
		1e	2e	3e	4e	5j	6j	
Rater	1	1	1	1	0	1	1	83%
	2	1	1	1	0	1	1	83%
	3	1	1	1	1	1	1	100%
	4	1	1	1	0	1	1	83%
	6	1	1	1	0	0	1	67%
	7	1	1	0	0	0	1	50%
	8	1	1	1	0	1	1	83%
	9	1	0	1	0	1	1	67%
	10	1	1	1	0	1	1	83%
	11	1	1	1	0	1	1	83%
	12	1	1	1	0	1	1	83%
	13	1	1	0	1	0	1	67%
	14	1	1	1	0	1	1	83%
	15	1	1	1	0	1	1	83%
	16	1	1	1	0	1	1	83%
	17	1	1	1	0	1	1	83%
	19	1	1	1	0	1	1	83%
	20	1	1	1	1	1	1	100%
Ergebnis Aufgabe %		100%	94%	89%	17%	83%	100%	

Schaubild 3 zeigt die Ergebnisse zu Aufgabe 2 – Produktion nach der ersten Bewertungsrunde, **vor** der Diskussion. Die Beispiele 1 und 6 wurden von allen übereinstimmend auf Niveau A2 eingestuft, Beispiel 4 von 15 Juror/-innen unter Niveau A2. Bei den Beispielen 2, 3 und 5 gab es jeweils eine/-n, zwei bzw. drei von 18 Juror/-innen, die diese Leistungen unter dem Niveau A2 einstuften.

Schaubild 4

Aufgabe 2 - Produktion		Beispiel						Ergebnis Rater %
		1e	2e	3e	4e	5j	6j	
Rater	1	1	1	1	0	1	1	83%
	2	1	1	1	0	1	1	83%
	3	1	1	1	0	1	1	83%
	4	1	1	1	0	1	1	83%
	6	1	1	1	0	1	1	83%
	7	1	1	0	0	0	1	50%
	8	1	1	1	0	1	1	83%
	9	1	1	1	0	1	1	83%
	10	1	1	1	0	1	1	83%
	11	1	1	1	0	1	1	83%
	12	1	1	1	0	1	1	83%
	13	1	1	1	0	1	1	83%
	14	1	1	1	0	1	1	83%
	15	1	1	1	0	1	1	83%
	16	1	1	1	0	1	1	83%
	17	1	1	1	1	1	1	100%
	19	1	1	1	0	1	1	83%
	20	1	1	1	0	1	1	83%
Ergebnis Aufgabe %		100%	100%	94%	6%	94%	100%	

Schaubild 4 zeigt die Ergebnisse zu Aufgabe 2 – Produktion **nach** der Diskussion der Abweichungen und der zweiten Bewertungsrunde. Auch hier führte die Diskussion in Teilgruppen zu einer stärkeren Einheitlichkeit der Voten bei den Beispielen 2, 3, 4 und 5. Bei Beispiel 2 sind nun alle der Meinung, es handle sich um eine Leistung auf A2-Niveau. Bei Beispiel 3 und 5 sind 17 von 18 der Meinung, es handle sich um eine Leistung auf A2-Niveau. Bei Beispiel 4 war nur ein/-e Juror/-in der Meinung, es handle sich um eine Leistung auf A2-Niveau, für alle anderen ist es eine Leistung unter A2-Niveau

Schaubild 5

Aufgabe 3 - Interaktion		Beispiel						Ergebnis Rater %
		1e	2e	3e	4e	5j	6j	
Rater	1	1	1	1	1	1	1	100%
	2	1	1	1	0	1	1	83%
	3	1	1	1	1	1	1	100%
	4	1	1	1	0	1	1	83%
	6	1	1	1	0	1	1	83%
	7	1	1	1	1	1	1	100%
	8	1	1	0	0	1	1	67%
	9	1	0	0	0	1	0	33%
	10	1	1	1	0	0	1	67%
	11	1	1	0	1	1	1	83%
	12	1	1	1	0	1	1	83%
	13	1	1	1	1	1	1	100%
	14	1	1	1	1	1	0	83%
	15	1	1	1	1	1	1	100%
	16	1	1	1	1	1	1	100%
	17	1	1	1	1	1	1	100%
	19	1	1	1	1	1	1	100%
	20	1	1	1	1	1	1	100%
Ergebnis Aufgabe %		100%	94%	83%	61%	94%	89%	

Schaubild 5 zeigt die Ergebnisse zu Aufgabe 3 – Interaktion nach der ersten Bewertungsrunde und **vor** der Diskussion. Das Beispiel 1 wurde von allen übereinstimmend auf Niveau A2 eingestuft. Bei den Beispielen 2 und 5 gab es jeweils eine/-n von 18 Ratern, die/der diese Leistungen unter dem Niveau A2 einstufte. Beispiel 6 stuften zwei Juror/-innen unter A2-Niveau ein. Bei Beispiel 3 entschieden drei Rater, die Leistung sei unter A2-Niveau. Bei Beispiel 4 waren sieben von 18 Juror/-innen der Meinung, die Leistung sei unter A2-Niveau.

Schaubild 6

Aufgabe 3 - Interaktion		Beispiel						Ergebnis Rater %
		1e	2e	3e	4e	5j	6j	
Rater	1	1	1	1	1	1	1	100%
	2	1	1	1	1	1	1	100%
	3	1	1	1	1	1	1	100%
	4	1	1	1	1	1	1	100%
	6	1	1	1	1	1	1	100%
	7	1	1	1	1	1	1	100%
	8	1	1	1	1	1	1	100%
	9	1	1	1	1	1	1	100%
	10	1	1	1	0	0	1	67%
	11	1	1	0	1	1	1	83%
	12	1	1	1	0	1	1	83%
	13	1	1	1	0	1	1	83%
	14	1	1	1	1	1	1	100%
	15	1	1	1	1	1	1	100%
	16	1	1	1	1	1	1	100%
	17	1	1	1	1	1	1	100%
	19	1	1	1	1	1	1	100%
	20	1	1	1	1	1	1	100%
Ergebnis Aufgabe %		100%	100%	94%	83%	94%	100%	

Schaubild 6 zeigt die Ergebnisse zu Aufgabe 3 – Interaktion **nach** der Diskussion der Abweichungen und der zweiten Bewertungsrunde. Bei Beispiel 1, 2, 3, 5 und 6 sind nun fast alle einstimmig der Meinung, die Leistungen seien auf A2-Niveau. Bei Beispiel 4 sind 15 von 18 Juror/-innen der Meinung, die Leistung sei auf A2-Niveau.

2.4.4 Statistische Analysen

Die statistischen Analysen mit dem Programm FACETS erbrachten Erkenntnisse über die Leistungsbeispiele (Cand), Jurorinnen und Juroren (rater), Teile (part) und Aufgaben (criteria). Die Ergebnisse wurden mit der Maßeinheit logits (Measr) dargestellt.

Schaubild 7

Einstufung der Teilnehmerleistungen, Rater, Aufgaben, Runde 1

Das Leistungsbeispiel 1e mit dem höchsten Wert hat +4 logits. Dieses Beispiel wurde von den meisten auf A2 eingestuft. Das Leistungsbeispiel 4e mit dem niedrigsten Wert hat –0,8 logits. Dieses Beispiel wurde von den meisten als unterhalb A2 eingestuft. Die Rater 9, 10 und 8 waren strenger als der Durchschnitt. Die Rater 20 und 3 waren die mildesten. Die anderen teilen sich in drei Gruppen zwischen –1,5 bis 0,5 logits. Aufgabe 2 zur Produktion wurde als schwieriger eingestuft als die Aufgaben 3 und 1 zur Interaktion.

Measr	Cand	rater	status	part	criteria
4	1e				
	6j				
	2e				
3		r9			
	5j				
	3e				
2					
		r10 r8			
1					
					Produktion
		r11 r14 r6 r7		Teil 2	
				Teil 3	
0			Runde 1		
		r12 r13 r19 r2			
	4e				Interaktion
				Teil 1	
-1					
		r1 r15 r16 r17			
-2		r20 r3			

Schaubild 8

Nach der Diskussion und dem zweiten Sehen veränderte sich die Bandbreite auf +7 bis −2 logits. Das bedeutet, dass die Leistungen im oberen Bereich nun als deutlich besser eingestuft wurden als nach der ersten Runde.

Das Leistungsbeispiel 4e, das auch nach Runde 1 den niedrigsten logit-Wert erhielt, wurde nach der zweiten Runde mit circa +0,5 logits etwas höher eingestuft. Es wurde damit gerade noch auf dem A2-Niveau gesehen. Ein Vergleich der beiden Schaubilder erlaubt es, die Bewegungen einzelner Jurorinnen und Juroren bezüglich Milde bzw. Strenge ihrer Bewertung nachzuverfolgen. Rater 10 gehört weiterhin zu den strengen. Rater 17 ist in der zweiten Runde der mildeste. Die Mehrheit der Juror/-innen liegt nun bei +1 und −2 logits. Sie haben sich nach der Diskussion aneinander angeglichen und bewerteten milder.

Measr	Cand	rater	status	part	criteria
7	1e 2e 6j				
	3e				
6					
5					
	5j				
4					
3					
		r10 r12 r7			
2					
				Teil 2	
		r11 r13 r19 r8			Produktion
1					
	4e				
0			Runde 2		
				Teil 3	
-1					
		r1 r14 r15 r16 r2 r20 r3 r4 r6 r9		Teil 1	Interaktion
-2		r17			

Schaubild 9

Intra-Rater-Zuverlässigkeit

Analysiert wurde zunächst die Differenz zwischen den in Runde 1 und Runde 2 abgegebenen Werten (0 oder 1). Mithilfe von Alpha (Spalte 2) wurde bestimmt, wie die beiden pro Aufgabe abgegeben Werte korrelierten. Wer in beiden Runden dieselbe Anzahl an 0- und 1-Werten vergab, erhielt den Wert 1.000 für komplette Übereinstimmung. Denselben Wert erhielt, wer zweimal identische Werte vergab. Mit dem Wert Kappa (Spalte 3) wird ebenfalls gezeigt, in welchem Maße die Werte von Runde 1 und 2 korrelieren. Wer in beiden Runden dieselben Werte vergab (siehe r1 und r7), erhielt einen Kappa-Wert 1.000. Der Wert in Spalte 4, Pearson, zeigt, wer konsistent seine Werte änderte. Im Schaubild markiert sind Rater mit geringer Zuverlässigkeit, geringer Übereinstimmung und geringer Konsistenz in ihren Bewertungen. Im Vergleich zu der Gruppe Schreiben zeigen sich in dieser Gruppe mehr hervorgehobene Werte. Jedoch zeigt nur eine von zwanzig Personen in Bezug auf alle drei Werte der Zuverlässigkeit, Übereinstimmung und Konsistenz Auffälligkeiten. Dies unterstreicht die Glaubwürdigkeit der im Benchmarking erzielten Ergebnisse.

rater	Alpha/ICC	Kappa	Pearson
r1	1.000	1.000	1.000
r2	0.790	0.686	0.640
r3*	N/A	0.000	N/A
r4	0.790	0.686	0.640
r6	0.652	0.542	0.455
r7	1.000	1.000	1.000
r8	0.778	0.661	0.609
r9	0.342	0.271	0.137
r10	0.908	0.837	0.824
r11	0.876	0.791	0.769
r12	0.876	0.791	0.769
r13	0.609	0.438	0.438
r14	0.652	0.542	0.455
r15	1.000	1.000	1.000
r16	1.000	1.000	1.000
r17*	N/A	0.000	N/A
r19	1.000	1.000	1.000
r20*	N/A	0.944	0.944

Schaubild 10

Inter-Rater-Zuverlässigkeit

Verglichen wurde die erwartete Übereinstimmung zwischen Juror/-innen mit beobachteten Übereinstimmungen, bezogen auf 2754 mögliche Übereinstimmungen in Runde 1 und 2754 in Runde 2.

	erwartet	**beobachtet**
Runde 1	84.0%	84.9%
Runde 2	92.2%	93.6%

Die Nähe zwischen erwarteter und beobachteter Übereinstimmung ist ein Indikator für eine sehr gute Inter-Rater-Zuverlässigkeit. Im vorliegenden Daten-Set war weniger als 1 Prozent der Beobachtungen in der Runde 1 unerwartet, in der 2. Runde etwas mehr als 1 Prozent. Die beteiligten Juror/-innen können somit als sehr zuverlässig bezeichnet werden.

Schaubild 11

Vergleich Benchmarking und Trainingsmaterialien

Beispiel	**Aufgabe 1**		**Aufgabe 2**		**Aufgabe 3**	
	Benchmarking	Training	Benchmarking	Training	Benchmarking	Training
Mariam	A2 100%	A2	A2 100%	A2	A2 100%	A2
Paula	A2 100%	A2	A2 94%	A2	A2 100%	A2
Simone	A2 100%	A2	A2 89%	A2	A2 94%	A2
Beka	A2 83%	A2	Unter A2 17%	Unter A2	A2 83%	A2
Liz	A2 100%	A2	A2 83%	A2	A2 94%	A2
Dyah	A2 100%	A2	A2 100%	A2	A2 100%	A2

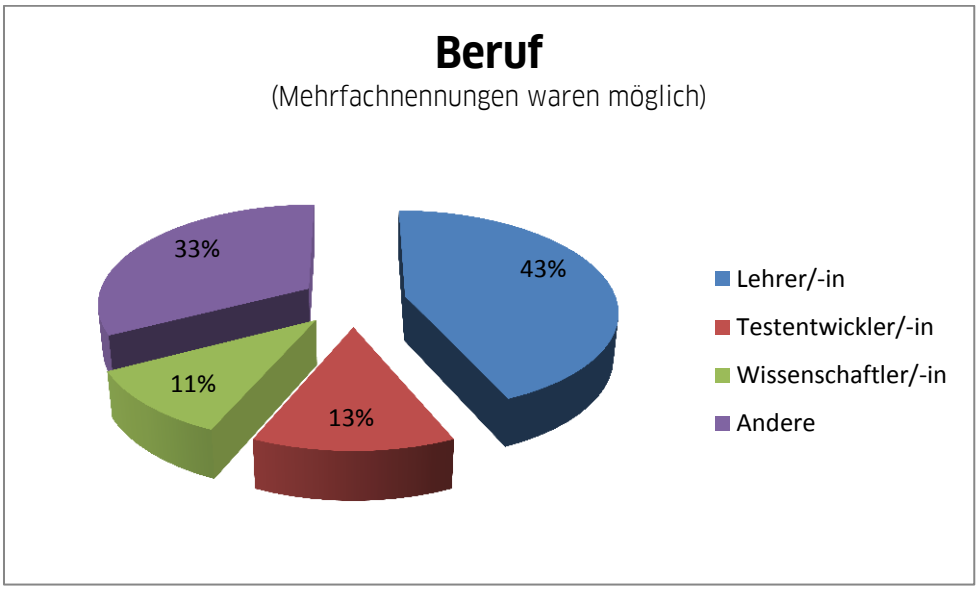
Schaubild 11 geht auf die Verbindung zwischen der Niveaueinstufung des Benchmarkings und den Ergebnissen der Bewertung durch Anwendung der für die Prüfung entwickelten Bewertungskriterien ein. Diese Ergebnisse finden sich in dem *Trainingsmaterial für Prüfende*. Die Gegenüberstellung zeigt, dass die Juror/-innen des Benchmarkings zu denselben Schlüssen gelangten wie die Bewertenden, die die Leistungsbeispiele im Trainingsmaterial bewerteten.

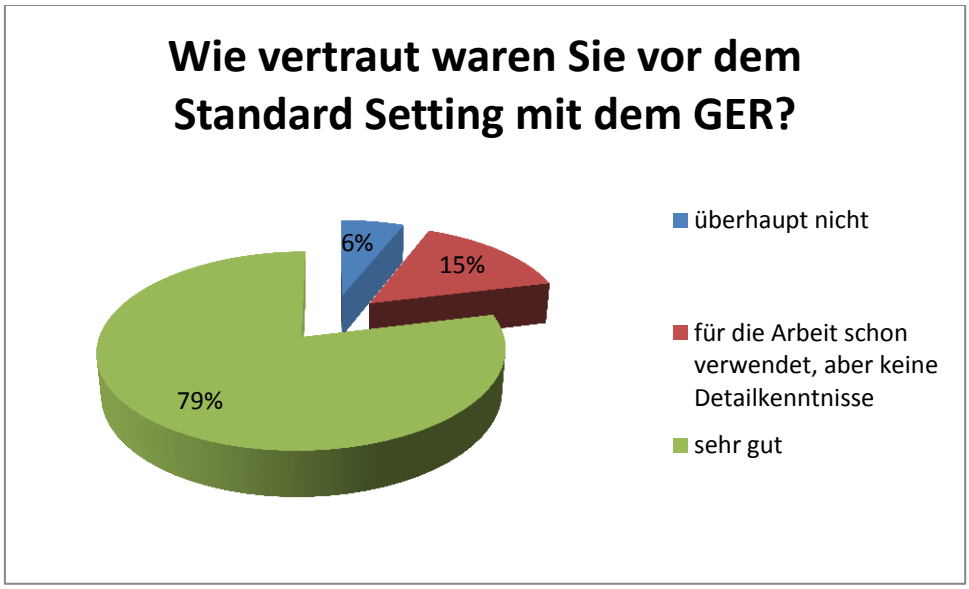
3 Evaluation der Veranstaltung

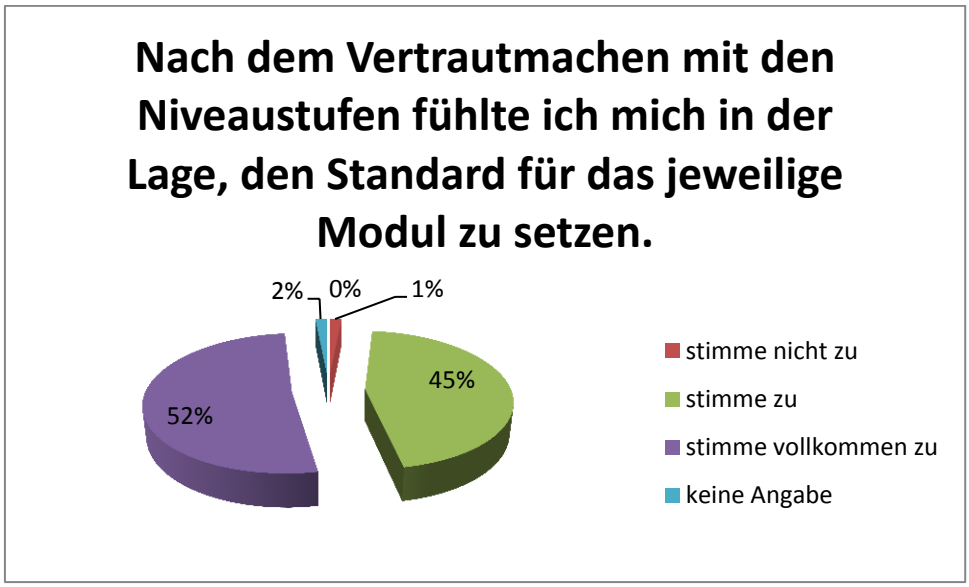
Die Teilnehmenden erhielten am Ende der Veranstaltung einen Evaluationsbogen. Sie zogen ein positives Fazit aus der Veranstaltung:

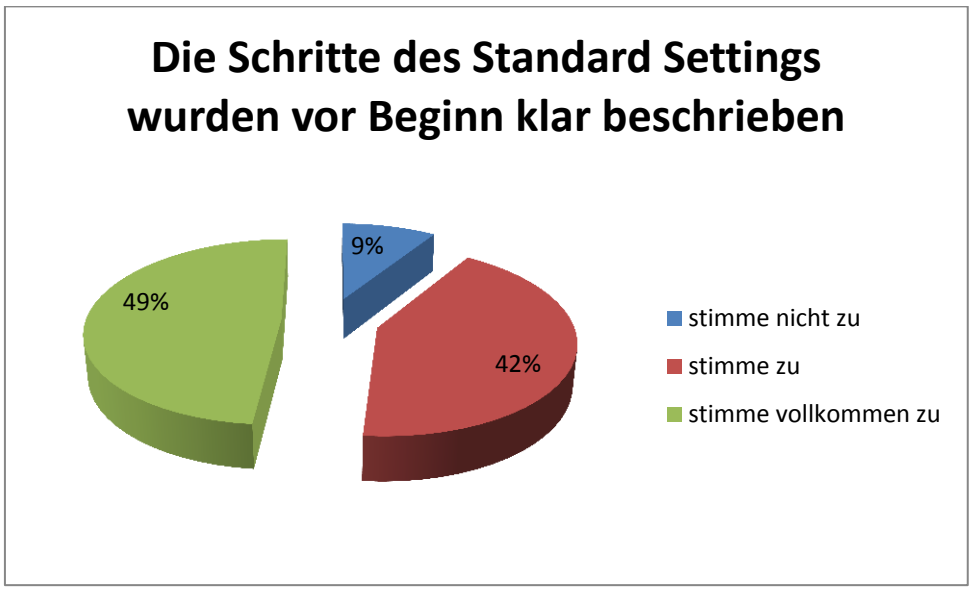
- *Danke, dass ich teilnehmen durfte, es ist eine große Unterstützung meiner Verlagsarbeit.*
- *Vielen Dank für die gute Organisation und nette Atmosphäre!*
- *Danke für die interessante Tagung. Es wäre interessant gewesen, alle 4 Fertigkeiten zu bewerten – doch dafür reicht die Zeit nicht.*
- *Sehr interessante Veranstaltung. Interessant auch, dass Bewertungen trotz Kriterienrasters „aus dem Bauch" getroffen werden und sprich immer interpretierbar bleibt.*

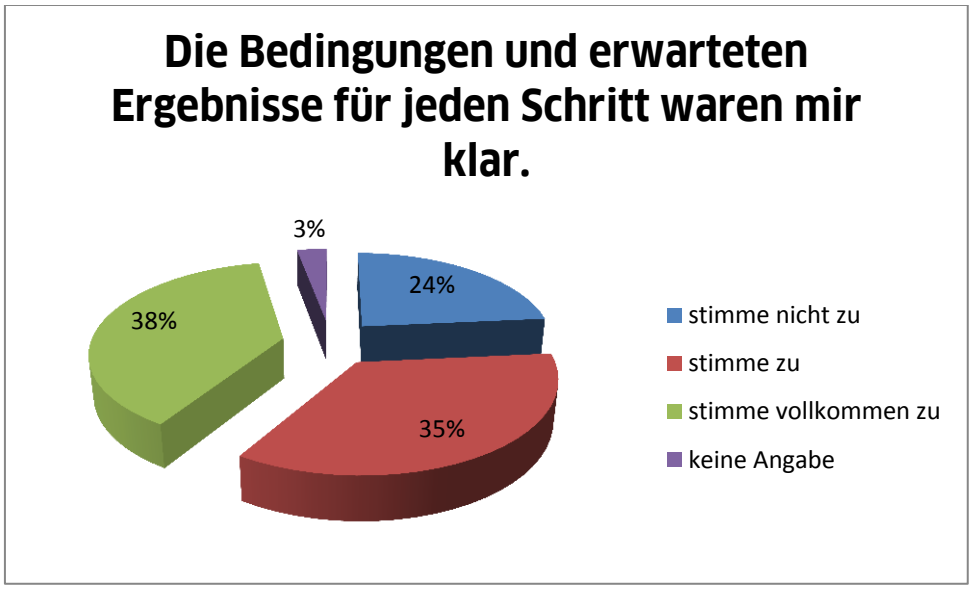
Die Teilnehmenden brachten teilweise auch Kritik zum Ausdruck. Es wurde unter anderem kritisiert, dass die Texte des Hörverstehens nicht lange bzw. oft genug vorgespielt wurden. Des Weiteren sollte im Item-Booklet darauf hingewiesen werden, ob der Text zu dem jeweiligen Item ein- oder zweimal gehört wird. Manche Teilnehmende wünschten sich außerdem mehr Zeit für die Diskussionen in den Gruppen und fanden die Aufgabenstellungen nicht immer klar formuliert. Dies sind wichtige Hinweise, die wir bei der Vorbereitung des nächsten Standard Settings und Benchmarkings berücksichtigen werden. Nachfolgend die quantitativen Ergebnisse:

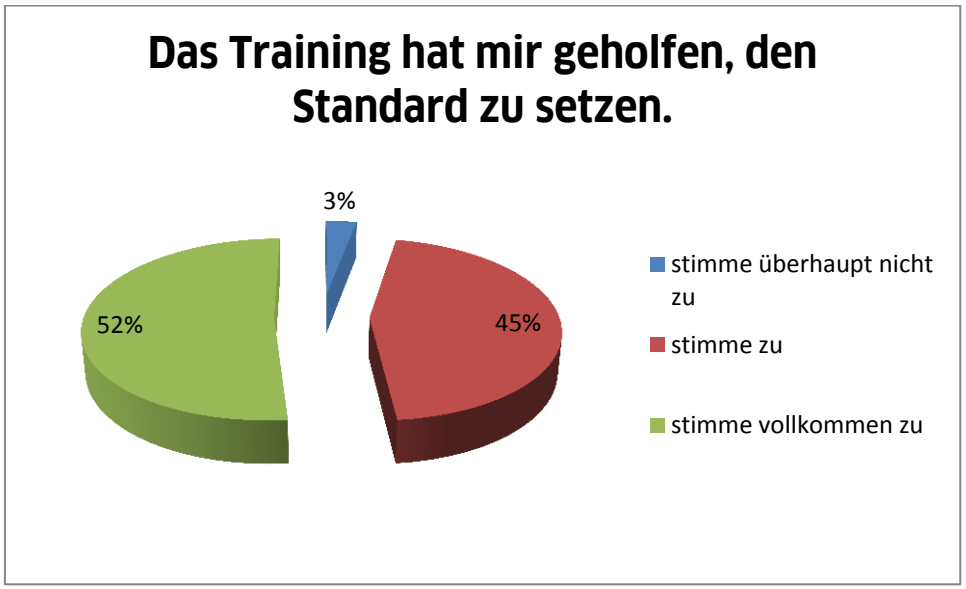


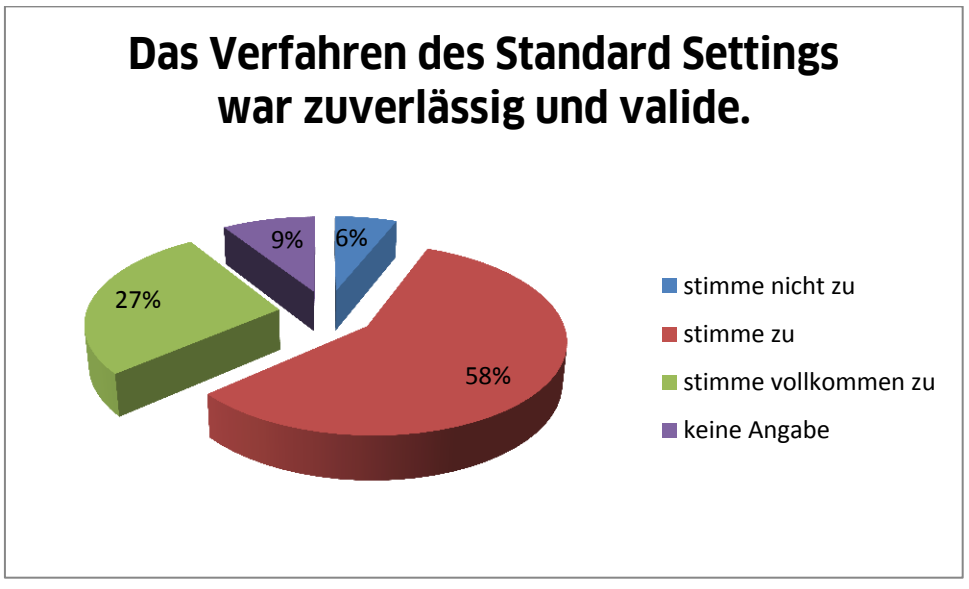

Wie vertraut waren Sie vor dem Standard Setting mit dem GER?
6%
15%
79%
überhaupt nicht
für die Arbeit schon verwendet, aber keine Detailkenntnisse
sehr gut

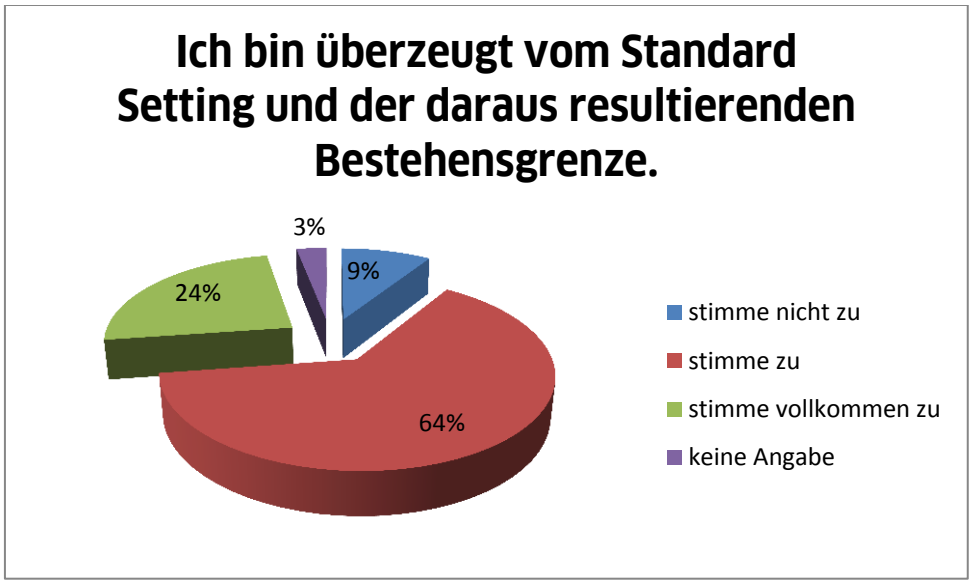

Nach dem Vertrautmachen mit den Niveaustufen fühlte ich mich in der Lage, den Standard für das jeweilige Modul zu setzen.
2%
0%
1%
45%
52%
stimme nicht zu
stimme zu
stimme vollkommen zu
keine Angabe

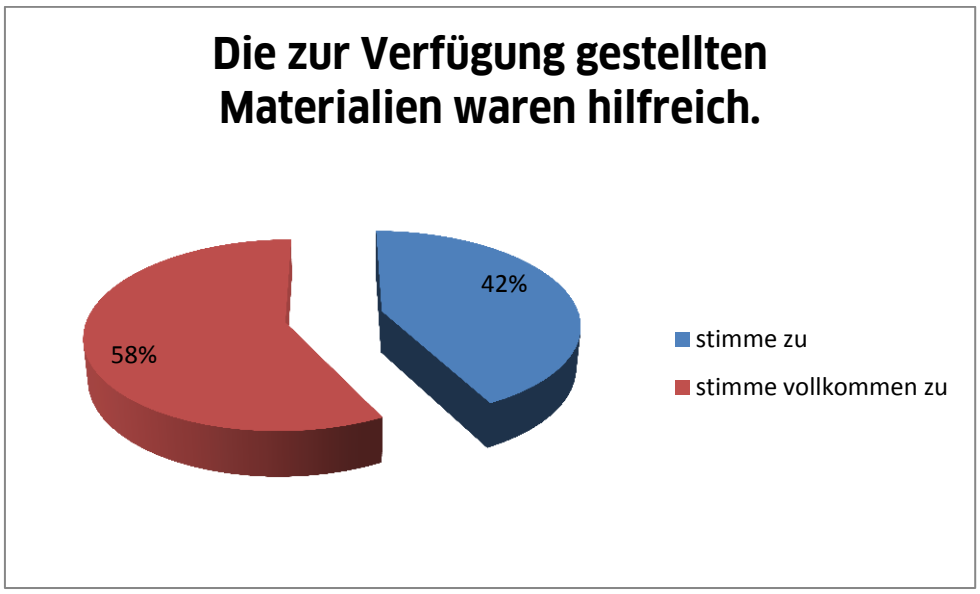

Die Schritte des Standard Settings wurden vor Beginn klar beschrieben
9%
49%
42%
stimme nicht zu
stimme zu
stimme vollkommen zu


Die Bedingungen und erwarteten Ergebnisse für jeden Schritt waren mir klar.
3%
24%
38%
35%
stimme nicht zu
stimme zu
stimme vollkommen zu
keine Angabe


Das Training hat mir geholfen, den Standard zu setzen.
3%
52%
45%
stimme überhaupt nicht zu
stimme zu
stimme vollkommen zu


Das Verfahren des Standard Settings war zuverlässig und valide.
9%
6%
27%
58%
stimme nicht zu
stimme zu
stimme vollkommen zu
keine Angabe


Ich bin überzeugt vom Standard Setting und der daraus resultierenden Bestehensgrenze.
3%
9%
24%
64%
stimme nicht zu
stimme zu
stimme vollkommen zu
keine Angabe


Die zur Verfügung gestellten Materialien waren hilfreich.
42%
58%
stimme zu
stimme vollkommen zu

4 Bibliografie

Association of Language Testers in Europe (ALTE) (2007): *Minimum standards for establishing quality profiles in ALTE examinations.* [Online.http://www.alte.org/attachments/files/minimum_standards.pdf. 11.04.2007].

Bachman, L.; Palmer, D. (2010): *Language Assessment in Practice*. Oxford: Oxford University Press (= Applied Linguistics).

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

Centre Internationale des Études Pedagogiques (2008): *Mündliche Leistungen: Beispiele für die 6 Niveaustufen des Gemeinsamen europäischen Referenzrahmens für Sprachen.* DVD. [Online: www.ciep.fr/publi_evalcert]

Europarat (2005): *Relating Language Examinations to the Common European Framework of References for Languages: Learning, Teaching, Assessment. Reading and Listening Items and Tasks: Pilot Samples illustrating the common reference levels in English, French, German, Italian and Spanish.* CD-ROM. Strasbourg: Council of Europe. [Online: http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Key_reference/exampleswriting_EN.pdf]

Europarat (2009): *Relating Language Examinations to the Common European Framework of References for languages: Learning, Teaching, Assessment. A manual*. Strasbourg: Council of Europe.

Europarat / ALTE (2011): *Manual for Language Test Development and Examining – For use with the CEFR.* Strasbourg: Council of Europe.

European Association of Language Testing and Assessment (EALTA)(2006): *Guidelines for Good Practice in Language Testing and Assessment* (Adopted 20th May 2006). [Online: http://www.ealta.eu.org/guidelines.htm. 01.03.2013].

Europarat (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen.* Berlin: Langenscheidt.

Fasoglio, D.; Beeker, A.; de Jong, K.; Keuning, J: Van Til, A. (2015): *CEFR-level writing skills for English, German and French*. Enschede: SLO (Netherlands Institute for Curriculum Development).

Feskens, R.; Keuning, J: van Til, A.; Verheyen, R. (2015) Performance Standards for the CEFR in Dutch secondary education. An international standard setting study. Cito: Arnhem.

Glaboniat, M.; Perlmann-Balme, M.; Studer, T. (2013): Zertifikat B1 Deutschprüfung für Jugendliche und Erwachsene: Standard Setting. Ein Arbeitsbericht. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, *18*, 72–75. [Online: http://zif.spz.tu-darmstadt.de/jg-18-1/beitrag/Glaboniat_Perlmann-Balme_Studer.pdf]

Glaboniat, M.; Müller, M.; Rusch, P.; Schmitz, H.; Wertenschlag, L. (2005): *Profile deutsch. A1 – C2 (Version 2.0).* Berlin: Langenscheidt.

Hennemann, D.; Perlmann-Balme, M.; Stelter, C (2015): Goethe-*Zertifikat A2. Deutschprüfung für Jugendliche und Erwachsene. Prüfungsziele, Testbeschreibung*. München: Hueber.

Kantarcıoğlu, E.; Papageorgiou, S. (2011): Benchmarking and standards in language tests. In: O'Sullivan, B. (Hrsg.): *Language testing. Theories and practices*. New York: Palgrave, S. 94–110.

Karantonis, A.; Sireci, S. G. (2006): The Bookmark Standard Setting Method: A Literature Review. In: *Educational Measurement: Issues and Practice 25*, 4–12.

Kecker, G. (2010): *Validierung von Sprachprüfungen. Die Zuordnung des TestDaF zum Gemeinsamen europäischen Referenzrahmen für Sprachen.* Frankfurt: Peter Lang.

Kenyon, D. (2013): Standard Setting on Language Tests. In: Chapelle, Carol A. (Hrsg.), *The Encyclopedia of Applied Linguistics*. Blackwell, 1-5. [Online: http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal1113/pdf].

Weir, C. J. (2005): *Language Testing and Validation: An evidence-based approach.* Houndgrave, Hampshire, UK: Palgrave-Macmillan.

Zeidler, B. (2016): Getting to know the minimally competent person. In: *Studies in Language Testing 44: Language Assessment for Multilingualism*: 251-269.

5 Anlagen

GOETHE-ZERTIFIKAT A2
STANDARD SETTING UND BENCHMARKING

München, 18. - 19. Januar 2016

PROGRAMM

Montag, 18.01.

	Plenum A080
14.00	Dr. Heike Uhlig: Begrüßung
	Johannes Gerbes: Die neue A2 Prüfung: Zahlen und Ziele
14.15	Claudia Stelter: Das Format des *Goethe-Zertifikat A2*
14.30	Dr. Michaela Perlmann-Balme: Methodik Standard Setting und Benchmarking
14.50	Dr. Doris Hennemann: Vertrautmachen mit Niveaustufe A2
15.15	**Kaffeepause**
	Gruppe **Lesen** Raum A113, Leitung: Doris Hennemann
	Gruppe **Hören** Raum A080, Leitung: Ekaterini Karamichali
15.40	Vertrautmachen mit Deskriptoren zu Fertigkeiten
16.00	Kennenlernen kalibrierter Testaufgaben des Europarats
16.15	Kennenlernen der Prüfungsteile *Goethe-Zertifikat A2*
16.30	Standard Setting Einzelarbeit mit Testaufgaben
17.15	Erstes Votum, Vorstellung der Ergebnisse
	Diskussion in Teilgruppen, Zweites Votum
18.00	Ende der Veranstaltung Tag 1
18.30	Gemeinsames Abendessen

Dienstag, 19.01.

	Gruppe **Schreiben** Raum A113, Leitung: Michaela Perlmann-Balme
	Gruppe **Sprechen** Raum A080, Leitung: Claudia Stelter
09.00	Vertrautmachen mit Deskriptoren zu Fertigkeiten
09.15	Kennenlernen kalibrierter Leistungsbeispiele des Europarats
09.30	Kennenlernen der Prüfungsteile *Goethe-Zertifikat A2*
09.45	Benchmarking Einzelarbeit mit Leistungsbeispielen, Runde 1
10.30	**Kaffeepause**
11.00	Vorstellung der Ergebnisse von Runde 1
	Diskussion in Teilgruppen, Sprechen: Zweites Votum
11.30	Einzelarbeit mit Leistungsbeispielen, Runde 2
	Diskussion in Teilgruppen, Zweites Votum
12.30	**Mittagessen**
	Plenum A080
13.30	Vorstellung der Ergebnisse aller Gruppen
14.15	Feedback, Evaluation
14.45	Ende der Veranstaltung

GOETHE INSTITUT

Sprache. Kultur. Deutschland.

GOETHE-ZERTIFIKAT A2
STANDARD SETTING UND BENCHMARKING

München, 18. - 19. Januar 2016

TEILNEHMERINNEN UND TEILNEHMER

Arras, Ulrike	Freie Universität Bozen/TestDaF-Institut
Beiser-Kolb, Vera	Landesverband der VHS Saarland
Bieber-Reynartz, Gudula	Münchner Volkshochschule
Breithaupt, Dominik	Sprachen & Dolmetscher Institut München
Bröcker, Kirsten	Landesverband der VHS Sachsen-Anhalt
Comparato, Giulia	Klett-Langenscheidt Verlag
Demmig, Silvia	Friedrich-Schiller-Universität Jena
Dengler, Stefanie	Goethe-Institut Zentrale
Göbels, Armin	Goethe-Institut Berlin
Heydenreich, Katharina	Deutsch perfekt/Spotlight Verlag
Hilger, Corinna	Cornelsen Schulverlage
Hoischen, Ina	Goethe-Institut Zentrale
Jacobs, Silke	Goethe-Institut Düsseldorf
Kellermann, Marco	Goethe-Institut Mannheim
Kirchner, Hildegard	Goethe-Institut Freiburg
Krüger, Tanja	Goethe-Institut Zentrale
Laub, Stefan	did deutsch-institut
Lloyd, Jane	Cambridge English Language Assessment
Loew, Sylke	Sprachenzentrum der Universität des Saarlandes
Loumiotis, Uta	Klett Hellas
Martyniuk, Waldemar	Jagiellonen-Universität in Krakau
Nimmrichter, Florian	Österreichisches Sprachdiplom
Plisch de Vega, Stefanie	Ernst Klett Sprachen
Rabl, Enikö	Ernst Klett Sprachen
Rapp, Arthur	Goethe-Institut Zentrale
Remus, Annerose	Klett-Langenscheidt Verlag
Schmid, Christiane	Sprachen & Dolmetscher Institut München
Shafer, Naomi	Université de Fribourg/Universität Freiburg
Sokolowski, Kathrin	Cornelsen Schulverlage
Staudigel, Irmingard	Landesverband der VHS Bayern
Suter Reich, Virginia	Zürcher Hochschule für Angewandte Wissenschaften
Tahy, Nora	Hueber Verlag
Terrasi-Haufe, Elisabetta	Ludwig-Maximilians-Universität München
Goossens, Dorrie	Central Instituut voor Toetsontwikkeling (Cito)
Widmann, Brigitte	Freie Universität Bozen
Widmer-Behr, Heike	Zürcher Hochschule für Angewandte Wissenschaften
Zimmermann, Sonja	TestDaF-Institut

GOETHE INSTITUT

Sprache. Kultur. Deutschland.