DAS MODULARE GOETHE-ZERTIFIKAT C1 STANDARD-SETTING UND BENCHMARKING

ERGEBNISSE DER ONLINE-VERANSTALTUNG AM 31.01. UND 02.02.2023

1. EINFÜHRUNG

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 aktuellen Stand der Testforschung anzupassen. Ziel einer Revision ist außerdem, die Prüfung möglichst genau am *Gemeinsamen europäischen Referenzrahmen für Sprachen* (GeR) und dessen Niveaubeschreibungen auszurichten. Das neue Goethe-Zertifikat C1 wird ab dem 1. Januar 2024 als modulare Prüfung an den Prüfungszentren des Goethe-Instituts eingesetzt. Der Modellsatz zur Prüfung findet sich als PDF zum Herunterladen auf der Homepage des Goethe-Instituts: www.goethe.de/vorbereitung

Die Zuordnung des neuen Goethe-Zertifikats C1 bzw. seiner Aufgaben zur GeR-Niveaustufe C1 ist ein wichtiger Bestandteil der Validierung im gesamten Entwicklungsprozess. Um die Ausrichtung am GeR nachzuweisen, wird das sogenannte *Standard-Setting und Benchmarking* mit externen Expert*innen durchgeführt und dokumentiert. Dabei geht es um den Nachweis, dass die Prüfungsanforderungen und die erhobenen Leistungsbeispiele der Teilnehmenden der Definition des angestrebten C1-Niveaus im Referenzrahmen entsprechen. Des Weiteren geht es um die Feststellung bzw. Bestätigung der in der Entwicklungsphase festgelegten Bestehensgrenze.

Die Online-Veranstaltung „Standard-Setting und Benchmarking Goethe-Zertifikat C1 modular (für Erwachsene)“ fand am 31.01. und 02.02.2023 mit insgesamt 38 Expert*innen für Deutsch als Fremdsprache aus verschiedenen Bildungsinstitutionen statt.
Die Erfahrungen sowie Ergebnisse des Standard-Settings und Benchmarkings zum neu entwickelten Goethe-Zertifikat C1 werden in Form dieses Kurzberichts veröffentlicht.

2. METHODEN DES STANDARD-SETTINGS UND BENCHMARKINGS

Der Terminus *Standard-Setting* bezeichnet ein strukturiertes Verfahren, um rezeptive Leistungen von Lernenden auf verbal definierte Niveaustufen wie diejenigen des GeR zu beziehen. Die Basis bilden dabei die durch Expert*innen individuell getroffenen oder in der Gruppe ausgehandelten Entscheidungen.[1] Gegenstand von Standard-Settings sind Testitems zu den rezeptiven Fertigkeiten mit ihren aus der Erprobung erhobenen statistischen Schwierigkeitswerten, die von einem Panel aus Fachleuten ausgehend von der zentralen Frage „Kann eine minimal kompetente Person auf C1-Niveau diese Items noch lösen?“ beurteilt werden. Ziel der Arbeitsgruppe zu den rezeptiven Modulen Lesen und Hören war es, den *Cut-off* (Bestehensgrenze) zu bestimmen, indem beurteilt wurde, wie viele Items des Moduls richtig gelöst werden müssen, um dieses zu bestehen. Dafür wurde festgestellt, wo die Expert*innen die Grenze zwischen „C1 erreicht" und „C1 noch nicht erreicht" sahen.
Bestimmt wurde diese Grenze unter Anwendung von zwei Methoden: In einer ersten Runde beurteilten die Expert*innen jedes einzelne Item mithilfe der modifizierten Angoff-Methode[2] sowie der Leitfrage, ob eine minimalkompetente Kandidatin bzw. ein minimalkompetenter Kandidat (MKK) das vorliegende Item eher richtig als falsch (= 1) oder eher falsch als richtig löst (= 0). Bei einer/einem

[1] Kecker (2010): 90 ff.; Cizek & Bunch (2007): 14 ff.; British Council et al. (2022): 47 f.
[2] Kecker (2010): 98 f.; British Council et al. (2022): 57.

MKK[3] handelt es sich um eine Person mit einer Kompetenz am unteren Rand des angestrebten C1-Niveaus.
In der zweiten Runde kam die Bookmark-Methode[4] zum Einsatz. Bei diesem Verfahren wurden die Items des Moduls Lesen bzw. Hören den Expert*innen in der Reihenfolge ihrer Schwierigkeit präsentiert, beginnend mit dem leichtesten Item. Die Schwierigkeitswerte basieren auf der Rasch-Analyse der Rückläufe aus Erprobungen mit bis zu 300 Teilnehmenden und werden in der Maßeinheit Logits angegeben. Die Expert*innen setzten dort ihr Bookmark – basierend auf ihrer Einschätzung aus Runde 1 –, wo sie die Bestehensgrenze für das C1-Niveau sahen.

Für die Beurteilung produktiver schriftlicher und mündlicher Leistungsbeispiele bearbeitete das Panel aus Fachleuten eine möglichst große Zahl an Leistungsbeispielen von Teilnehmenden. Für diesen Arbeitsschritt hat sich der Begriff *Benchmarking* etabliert. Ziel der Arbeitsgruppe zu den produktiven Modulen Schreiben und Sprechen war es nachzuweisen, dass sich die auf Basis der Aufgaben des Modelltests C1 erhobenen Leistungsbeispiele von Teilnehmenden mit der Definition des C1-Niveaus im GeR decken. Ein weiteres Ziel bestand darin, eine Reihe von Referenzleistungen zu erhalten, die von Expert*innen klar auf dem C1-Niveau verortet wurden. Beurteilt wurden die Leistungsbeispiele unter Anwendung der Benchmarking-Methode[5].

Beiden Verfahren (Standard-Setting und Benchmarking) waren folgende Phasen der Familiarisierung vorgeschaltet:

1. **Phase**: Die Expert*innen befassten sich vor der Live-Sitzung mithilfe von zwei Arbeitsblättern mit den Deskriptoren des C1-Niveaus, insbesondere in Abgrenzung zu den benachbarten Niveaustufen B2 und C2. Dies diente als Vorbereitung auf die Diskussion, die während der Veranstaltung in Kleingruppen und im Plenum geführt wurde.
2. **Phase**: Die Expert*innen füllten in Einzelarbeit eine Umfrage zum/zur minimalkompetenten Kandidaten/Kandidatin (MKK) aus. Dabei wurden zentrale Deskriptoren präsentiert, für die die Expert innen den Grad ihrer Zustimmung (stimme voll zu – stimme zu – stimme nicht zu) zu der Aussage „Diese Kompetenz hat ein/eine MKK auf C1-Niveau schon" angeben mussten. Ziel war die Auseinandersetzung mit einer/einem MKK auf C1-Niveau, um ein gemeinsames Verständnis zu entwickeln, das für die Weiterarbeit essenziell war.
3. **Phase**: Die Expert*innen bewerteten kalibrierte Aufgaben/Leistungen des Europarats auf C1-Niveau sowie den angrenzenden Niveaustufen B2 und C2 und diskutierten ihre Einschätzung hinsichtlich der Merkmale des Niveaus C1.

Anschließend fand das Standard-Setting bzw. Benchmarking mit Aufgaben des Modellsatzes bzw. mit Leistungsbeispielen der Teilnehmenden zum modularen Goethe-Zertifikat C1 statt.
Die Standard-Setting- und Benchmarking-Konferenz zum modularen Goethe-Zertifikat C1 wurde als Online-Veranstaltung durchgeführt. Einschlägige Literatur[6] zur Durchführung des Standard-Settings

[3] Kecker (2010): 93.; Zeidler (2016): 252 ff.
[4] Kecker (2010): 99 f.; British Council et al. (2022): 60.
[5] Kecker (2010): 89 f.
[6] Katz & Tannenbaum (2014); British Council et al. (2022): 60; Kecker & Eckes (2021).

im Online-Format beschreibt beide Verfahren (digital vs. in Präsenz) als vergleichbar. Die Vergleichbarkeit bezieht sich dabei insbesondere auf die Zuverlässigkeit der Beurteilungen durch die Expert*innen.

3. ERGEBNISSE

Das Ziel des Standard-Settings für die Module Lesen und Hören war es, die Ausrichtung auf das Zielniveau C1 des GeR und die Bestehensgrenze von 60 % durch Expert*innen überprüfen zu lassen. Die Expert*innen bestätigten das Niveau der Prüfung und den in der Entwicklung festgelegten Cut-off mit hoher Übereinstimmung.
Ziel des Benchmarkings war es sicherzustellen, dass mit der Bearbeitung der Aufgaben des Moduls Schreiben bzw. Sprechen schriftliche bzw. mündliche Leistungen auf C1-Niveau von Teilnehmenden produziert werden können. Die Ergebnisse des Benchmarkings zeigen, dass dieses Ziel erreicht wurde.

Die Ergebnisse des Benchmarkings sind urteilsbasierter Natur. Daher ist es sinnvoll, eine Evaluation der Zuverlässigkeit der Expert*innen bzw. der Konsistenz ihres Bewertungsverhaltens vorzunehmen. Die statistischen Analysen hierzu wurden mit dem Programm FACETS erstellt und die Ergebnisse in der Maßeinheit Logits dargestellt. Mit Hilfe des Programms lassen sich über die bisherigen Aussagen hinaus Erkenntnisse zum Verhalten der Expert*innen gewinnen. Analysiert wurde dazu die Differenz zwischen den in Runde 1 und Runde 2 abgegebenen Werten.
Bei der Beurteilung der Schreibleistungen lassen sich so von 18 Expert*innen vier identifizieren, die statistisch gesehen ein inkonsistentes Bewertungsverhalten gezeigt haben.
Bei der Beurteilung der mündlichen Leistungen der Teilnehmenden lässt sich von 15 Expert*innen nur ein oder eine Person identifizieren, die statistisch gesehen ein inkonsistentes Bewertungsverhalten gezeigt hat.

4. FEEDBACK UND HINWEISE FÜR ZUKÜNFTIGE VERANSTALTUNGEN

Die Teilnehmenden hatten zu mehreren Zeitpunkten die Möglichkeit, Rückmeldung zur Veranstaltung zu geben. Eine Empfehlung für die zukünftige Durchführung solcher Veranstaltungen im Online-Format ist das Einplanen eines größeren zeitlichen Puffers. Die Zeitslots in den Arbeitsgruppen zu den einzelnen Aktivitäten waren so dicht geplant, dass es wenig Möglichkeiten gab, zeitlich umzudisponieren. Der enge Zeitplan wiederum war Resultat der Umstellung auf das Online-Format. Eine Möglichkeit der Entzerrung wäre das Einplanen einer zusätzlichen Pause, um die Arbeit in den Gruppen ggf. 30 bis maximal 60 Minuten länger gestalten zu können.
Nach Abschluss der Veranstaltung wurden die Teilnehmenden gebeten, an einer ausführlichen Online-Evaluation über das Tool *Inquery* teilzunehmen. Die Teilnehmenden meldeten ein hohes Maß an Zufriedenheit mit der Veranstaltung allgemein sowie dem Online-Verfahren des Standard-Settings und Benchmarkings zurück.

5. BIBLIOGRAFIE

Association of Language Testers in Europe (ALTE) (2007): *Minimum standards for establishing quality profiles in ALTE examinations.* [Online: http://www.coe.int/t/dg4/education/elp/elpreg/Source/Key_reference/exampleswriting_EN.pdf]

British Council, Ukalta, EALTA & ALTE (2022): *Aligning Language Education with the CEFR: A Handbook*. [Online: https://alte.org/resources/Documents/CEFR%20alignment%20handbook%20layout.pdf]

Cizek, G., Bunch M. (2007): *Standard Setting. A guide to establishing and evaluation performance standards on tests*. Thousand Oaks: Sage Publications.

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

Europarat (2020): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen. Begleitband*. Stuttgart: Klett Sprachen.

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. (DVD)

Katz, I. R., Tannenbaum, R. J. (2014): *Comparison of web-based and face-to-face standard setting using the Angoff Method.* In: Journal of Applied Testing Technology, 15(1), 1–17.

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

Tannenbaum, R. J., Cho, Y. (2014): *Critical Factors to Consider in Evaluating Standard-Setting Studies to Map Language Test Scores to Frameworks of Language Proficiency*. In: Language Assessment Quarterly, 11(3) 233–249.
[Online: http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal1113/pdf]

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

Goethe-Institut e. V.
Abteilung 40, Bereich 41 DaF-Prüfungen
Oskar-von-Miller-Ring 18
80333 München

pruefungen@goethe.de