Sonja Zimmermann, Daniela Marks

Integrierte Prüfungsleistungen mit dem GER beurteilen

Was die Kann-Beschreibungen nicht können

Abstract

Since its publication in 2001, the CEFR has become an internationally recognised framework for describing learners' language competence. Many test providers use the common reference levels and the illustrative descriptors for test development and rating purposes, although the scales of the CEFR were not meant to serve this purpose. For rating written or spoken performances, the very general descriptors of the CEFR have to be adapted to the specific assessment context. This is all the more crucial when it comes to the assessment of integrated tasks: both the transformation of the content and the language used in the source text have to be taken into account. However, the required cognitive operations and competences for this transformation are not fully represented in the current version of the CEFR. This paper highlights some theoretical implications of rating scale development for integrated tasks, and presents preliminary results of piloting task-relevant rating scales in an academic context.

1. Integrierte Aufgaben

Integrierte Schreib- und Sprechaufgaben werden zunehmend in kompetenzorientierten Sprachprüfungen eingesetzt (vgl. z.B. Knoch & Sitajalabhorn 2013; Eckes, Müller-Karabil & Zimmermann 2016). Darunter sind solche Aufgabenformate zu verstehen, in denen Prüfungsteilnehmer[1] einen längeren Input in Form eines Lese- und/oder Hörtextes (ggf. unter Einbezug von Grafiken) in einen produktiven Output verarbeiten müssen (Lee 2006; Plakans 2013). Die Attraktivität dieser Aufgabenformate für das Überprüfen der Schreib- und Sprechkompetenz – besonders im akademischen Kontext – liegt vor allem darin begründet, dass die Fähigkeit, Informationen aus unterschiedlichen Quellen in eigenen schriftlichen oder mündlichen Äußerungen zu verarbeiten, als sprachliche Voraussetzung dafür angesehen wird, erfolgreich am Wissenschaftsdiskurs teilhaben zu können (vgl. Dittmann et al. 2003; Ehlich & Steets 2000; Jakobs 1997). Ein wichtiger Schritt in diesem Verarbeitungsprozess ist das Zusammenfassen, d.h. die Reduktion des Rezipierten auf wesentliche Hauptaussagen und die anschließende Wiedergabe dieser Einzelaussagen in einem kohärenten Text (Kruse & Ruhmann 2003). Integrierte Schreib- oder Sprechaufgaben, die von den Teilnehmenden eine schriftliche oder mündliche Zusammenfassung eines gelesenen oder gehörten Textes verlangen, können somit zur Validität von Sprachprüfungen im Hochschulkontext beitragen, da sie diese Kompetenz weitgehend authentisch widerspiegeln (vgl. Cumming 2013).

Dennoch werden integrierte Prüfungsaufgaben in der Forschung unterschiedlich bewertet: Die Vermischung verschiedener Fertigkeiten – gelegentlich auch als „muddied measurement" (Weir 2005: 101) bezeichnet – wird besonders vor dem Hintergrund der Frage nach dem zugrundeliegenden Konstrukt kritisch gesehen. Denn es bleibt die Frage, welche Fertigkeit letztendlich die Leistung bestimmt. Daher sind für die Auswertung und Beurteilung integrierter Prüfungsleistungen passgenaue Beurteilungskriterien relevant (vgl. Chan, Inoue & Taylor 2015; Yu 2013).

1 Aus Gründen der sprachlichen Vereinfachung werden in diesem Beitrag Ausdrücke wie „Prüfungsteilnehmer", „Beurteiler", „Proband" usw. im generischen Sinne verwendet.

2. Beurteilen mit dem GER – Was die Kann-Beschreibungen nicht können

Legt man die Skalen des *Gemeinsamen Europäischen Referenzrahmens für Sprachen* (GER) für die Beurteilung produktiver Prüfungsleistungen zugrunde, so zeigt sich, dass sie nur bedingt für diesen Zweck geeignet sind. Dies liegt vor allem daran, dass die GER-Skalen in allererster Linie benutzerorientiert sind und daher sehr allgemeine und wenig spezifische Beschreibungen von Sprachkompetenz enthalten. Beurteilungsskalen dagegen erfordern besonders genaue Ausführungen von relevanten Aspekten der Schreib- bzw. Sprechkompetenz bezogen auf den jeweiligen Kontext (vgl. Harsch & Martin 2012). Die Notwendigkeit, die allgemeinen Kann-Beschreibungen des GER an den Prüfungskontext anzupassen, gilt umso mehr, wenn Prüfungsleistungen beurteilt werden, denen integrierte Aufgaben zugrunde liegen. Exemplarisch soll dies am Beispiel einer (schriftlichen oder mündlichen) Zusammenfassung erläutert werden: So gibt es im GER zwar eine Skala „Texte verarbeiten" (GER 2001: 98), die für die einzelnen Kompetenzniveaus beschreibt, in welcher Qualität Lerner unterschiedlich komplexe Texte zusammenfassen können. Die dafür erforderlichen kognitiven Operationen und sprachlichen Kompetenzen werden in der aktuellen Fassung des GER[2] jedoch nicht so umfassend ausgeführt, um daraus Deskriptoren für die Beurteilung integrierter Aufgaben abzuleiten, die sowohl die inhaltliche als auch die sprachliche Verarbeitung des Quellenmaterials widerspiegeln sollen (Knoch & Sitajalabhorn, 2013).

Im Rahmen eines Projekts am TestDaF-Institut zur Entwicklung von Beurteilungsskalen für integrierte Prüfungsaufgaben wurden daher zunächst die wesentlichen Charakteristika der geforderten Textsorte und die dafür erforderlichen Kompetenzen festgehalten: So sind zentrale Merkmale einer Zusammenfassung, dass die relevanten Informationen aus dem Originaltext komprimiert und korrekt darstellt werden und dabei auf unwichtige Details sowie auf zusätzliche Informationen, die nicht in diesem Text enthalten sind, verzichtet wird. Voraussetzung dafür ist, dass Prüfungsteilnehmer über eine mentale Repräsentation des rezipierten Textes verfügen. Erst dann ist es ihnen möglich, die Informationen durch ein breites Spektrum sprachlicher Mittel so verdichtet und eigenständig zusammenzufassen, dass der Leser bzw. Zuhörer über den Inhalt des Textes

2 Eine erweiterte Fassung der GER-Skalen, die die kognitiven Verarbeitungsprozesse für die Vermittlung von gelesenen und/oder gehörten Informationen stärker berücksichtigt, befindet sich gerade in der Erprobungsphase (vgl. Council of Europe 2016).

informiert wird.[3]

Tabelle 1 zeigt, welche relevanten Aspekte sich somit in Bezug auf die rezeptiven und produktiven Verarbeitungsprozesse des Ausgangstextes ergeben, und wie diese – neben weiteren Aspekten wie Korrektheit oder auch Aussprache und Intonation im Mündlichen – für die Bewertung einer schriftlichen oder mündlichen Zusammenfassung in den inhaltlichen und sprachlichen Kriterien abgebildet werden können:

Kriterium	Aspekte (Rezeption & Produktion)
Inhalt	relevante Informationen aus Ausgangstext enthalten korrekte, strukturierte und komprimierte Wiedergabe der Informationen
Sprache	Spektrum sprachlicher Mittel zur eigenständigen Formulierung lexikalische Breite, d.h. präzise Verwendung relevanter Begriffe (auch aus dem Ausgangstext)

Tabelle 1: Abbildung der Verarbeitungsprozesse in den Beurteilungskriterien für schriftliche oder mündliche Zusammenfassungen

3. Pilotierung der Skalen

Aufgabe

Im Verlaufe des Projekts wurden Bewertungsskalen für verschiedene integrierte Aufgaben entwickelt, u.a. auch für eine mündliche Zusammenfassung eines längeren schriftlichen Inputtextes (ca. 250 Wörter). Für das Lesen des Textes und die Vorbereitung der Antwort standen insgesamt 4 Minuten zur Verfügung; die Sprechzeit für die Zusammenfassung betrug 2 Minuten. Der Inputtext lag den Probanden nur in der Vorbereitungszeit vor, nicht mehr während der Sprechzeit. Die Antworten der Teilnehmenden wurden am PC aufgezeichnet.

Teilnehmer

Die Aufgabe wurde von insgesamt 110 Probanden bearbeitet. Dabei handelte es sich um Deutschlerner, die im Durchschnitt 25 Jahre alt waren,

3 Vgl. dazu die Ausführungen von Yu (2013) für das Englische sowie von Stezano Cotelo (2008) für das Deutsche.

etwas mehr als die Hälfte davon weiblich (57%). Die Teilnehmer kamen aus mehr als 30 Ländern, die beiden größten Teilnehmergruppen stammten aus China (21 Personen) und Syrien (11 Personen). Vor der Bearbeitung der integrierten Aufgabe wurden die Teilnehmenden gebeten, den Einstufungstest onSET Deutsch abzulegen, um ihre allgemeine Sprachkompetenz einschätzen zu können. Die unten stehende Tabelle zeigt die Leistungsverteilung der Probanden.

	onSET-Ergebnisse der Teilnehmer (in %)
C1 oder höher	3,7 %
B2.2	14,6 %
B2.1	35,4 %
B1	45,1%
A2 oder niedriger	1,2 %

Tabelle 2: Verteilung der Leistungsniveaus der Probanden für die integrierte Aufgabe

Beurteiler

Die Aufnahmen der Teilnehmerantworten wurden von insgesamt 10 erfahrenen Beurteilern (2 Männer, 8 Frauen) eingestuft. Der Kreis der Beurteiler setzte sich zusammen aus Experten im Bereich der Skalenentwicklung, aus Hochschuldozenten mit DaF-Hintergrund und Deutschlehrkräften mit mehrjähriger Erfahrung in der Leistungsbeurteilung von ausländischen Studienbewerbern.

Methodisches Design

Zur Einstufung der Leistungen lagen den Beurteilern analytische Beurteilungsskalen vor, die insgesamt fünf verschiedene Aspekte der Teilnehmerantwort betrachteten: drei sprachliche Aspekte (bezogen auf Grammatik, Wortschatz und Aussprache/Intonation) und zwei inhaltliche Aspekte (Zusammenfassung und Struktur, s.o.). Dabei spiegelten die beiden inhaltlichen Kriterien und das Wortschatzkriterium die besonderen Anforderungen der integrierten Aufgabe wider. Für alle fünf Kriterien

wurden Deskriptoren auf vier Leistungsstufen formuliert, die etwa den Stufen B1 bis C1.2 des GER entsprechen. Die Beurteiler waren aufgefordert, jede Teilnehmerleistung in jedem Kriterium einzustufen.

Von den 110 Teilnehmeraufnahmen wurden 100 von je zwei Beurteilern bewertet, die verbleibenden 10 Antworten nur einfach. Die Teilnehmerantworten wurden so auf die Beurteiler verteilt, dass sich für jeden einzelnen Beurteiler Überschneidungen mit vier anderen Beurteilern ergaben. Die resultierende konnektive Datenmatrix erlaubte es, die Einstufungen der Beurteiler miteinander zu vergleichen. Mithilfe der Multifacetten-Rasch-Analyse wurden die Strenge bzw. Milde sowie die Beurteilungskonsistenz ermittelt (vgl. Eckes 2015). Diese Analysen erlauben erste Rückschlüsse hinsichtlich der Handhabbarkeit der Bewertungskriterien. Die statistischen Auswertungen wurden ergänzt durch qualitatives Feedback der Beurteiler in Form von schriftlichen Rückmeldungen und Diskussionen.

Erste Ergebnisse

Erste Analysen zeigen, dass die Beurteiler die Bewertungskriterien und die Deskriptoren der Skalen bei der Bewertung gut anwenden konnten. Die Konsistenz in ihren Urteilen ist zufriedenstellend und auch die Strenge bzw. Milde der Beurteiler liegt in einem akzeptablen Bereich, der eine hinreichende Differenzierungsfähigkeit gewährleistet.

Analysiert man die Bewertungskriterien, die sich auf die besonderen Anforderungen der o.g. integrierten Aufgabe beziehen, so zeigt sich, dass die beiden *inhaltlichen* Kriterien (die Zusammenfassung des gelesenen Inputs und die Struktur der eigenen Äußerung) statistisch unauffällig waren. Die qualitativen Rückmeldungen der beteiligten Beurteiler zeigen jedoch, dass es z.T. Unsicherheiten in der Anwendung der inhaltlichen Kriterien gab. Insbesondere Teilnehmerleistungen, in denen lediglich die Hauptaussage des Textes wiedergegeben und ansonsten eigenes Wissen präsentiert wurde, waren mit den Deskriptoren nicht zu fassen und daher schwer zu beurteilen.

Nach den statistischen Analysen ist bei dieser Aufgabe vielmehr die Beurteilung *sprachlicher* Aspekte problematisch: Betrachtet man das Konsistenzmaß der Beurteilungen im Wortschatzkriterium, zeigt sich eine leichte Tendenz zu wenig Varianz in den Urteilen. Dies kann auf die schwache Teilnehmergruppe zurückzuführen sein, die kaum gute Leistungen hervorgebracht hat. Möglicherweise ist der Wert aber auch ein Indiz für den Überarbeitungsbedarf der Deskriptoren in diesem Krite-

rium. Die Beurteiler selbst meldeten eher Probleme mit dem Kriterium, das sich auf die Breite des grammatikalischen Spektrums bezog: Hier stellte sich die Frage, wie mit Teilnehmerantworten umzugehen ist, die für ihre Zusammenfassung die syntaktischen Strukturen des Inputtextes übernehmen und kaum selbstständig formulieren.

4. Diskussion und Ausblick

Bei der Bewertung der Ergebnisse aus der Pilotierung der Beurteilungsskalen muss berücksichtigt werden, dass mehrere Aspekte die Interpretation einschränken:

Insgesamt war die Gruppe der Teilnehmer sprachlich deutlich schwächer als ursprünglich angestrebt. Da die Schwierigkeit der Aufgabe im oberen B2- bis C1-Bereich zu verorten ist, sollte auch das Sprachniveau der Probanden in diesem Bereich liegen, damit sie sprachlich in der Lage sind, die Aufgabe wie intendiert zu bearbeiten. Tatsächlich erreichte aber nur etwa ein Sechstel von ihnen eine Einstufung auf B2.2 oder darüber. Hinzu kommt, dass die Teilnehmer den Aufgabentyp vorab nicht kannten und sich daher nicht vorbereiten konnten.

Die Ergebnisse sind auch mit Blick auf die Beurteiler mit Vorsicht zu interpretieren: Zwar wurden die Aufgabenanforderungen und die Bewertungskriterien vor der Bewertung mit den beteiligten Beurteilern besprochen, auf eine ausgiebige Schulung mit praktischer Anwendung der Bewertungskriterien und anschließende Diskussion musste jedoch aus Zeitmangel verzichtet werden. Die Beurteiler konnten in ihrer Arbeit auch nicht auf Kalibrierungsunterlagen mit *Benchmarks* und Erwartungshorizont zurückgreifen, die die Bewertungsarbeit zusätzlich standardisiert hätten.

Die vorliegenden Teilnehmerantworten sollen noch weiteren Analysen unterzogen werden. Von Interesse sind hier insbesondere linguistische Analysen, die zeigen sollen, in welchem Maße Teilnehmer sich an den sprachlichen Strukturen und am Wortschatz des Inputtextes bedienen bzw. umgekehrt inwieweit eigenständige Formulierungen erwartbar sind. Dabei muss zum einen unterschieden werden, ob ein (flüchtiger) gehörter oder ein gelesener Input vorliegt, und zum anderen, ob bei der Produktion ein schriftlicher oder mündlicher Text zu verfassen ist.

Nach der Auswertung der relevanten Daten ist eine Überarbeitung der Beurteilungsskalen erforderlich, um u.a. die einzelnen Kriterien deutlicher voneinander abzugrenzen und die Deskriptoren passgenauer zu formulieren. Der nächste Schritt ist eine weitere Erprobung mit neuen und

überarbeiteten Aufgaben sowie einer größeren Teilnehmergruppe auf höherem Sprachniveau.

Literatur

Chan, Sathena/Inoue, Chihiro/Taylor, Lynda (2015): Developing rubrics to assess the reading-into-writing skills: A case study. In: *Assessing Writing*, 26, S. 20-37.

Council of Europe (2016): CEFR Illustrative Descriptors. Extended Version. Pilot version for consultation. Language Policy Unit: Strasbourg.

Cumming, Alister (2013): Assessing integrated writing tasks for academic purposes. Promises and Perils. In: Language Assessment Quarterly, 10 (1), S. 1-8.

Dittmann, Jürgen/Geneuss, Katrin A./Nennstiel, Christoph/Quast, Nora A. (2003): Schreibprobleme im Studium. Eine empirische Untersuchung. In: Ehlich, Konrad/ Steets, Angelika (Hrsg.): Wissenschaftlich schreiben – lehren und lernen. Berlin/New York: De Gruyter, S. 155-185.

Eckes, Thomas (2015): Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments. Frankfurt/Main: Peter Lang.

Eckes, Thomas/Müller-Karabil, Anika/Zimmermann, Sonja (2016): Assessing writing. In: Tsagari, Dina/Banerjee, Jayanti (Hrsg.): Handbook of second language assessment. Boston, MA: De Gruyter, S. 147-164.

Ehlich, Konrad/Steets, Angelika (2000): Schreiben im Studium. In: Einsichten. Forschung an der LMU 2/2000, S. 47-50.

Europarat (2001): Gemeinsamer Europäischer Referenzrahmen für Sprachen. Lehren, Lernen, Beurteilen. Berlin: Langenscheidt.

Harsch, Claudia/Martin, Guido (2012): Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. In: Assessing Writing, 17 (4), S. 228-250.

Jakobs, Eva-Maria (1997): Lesen und Textproduzieren. Source reading als typisches Merkmal wissenschaftlicher Textproduktion. In: Jakobs, Eva-Maria/Knorr, Dagmar (Hrsg.): Schreiben in den Wissenschaften. Frankfurt/Main: Peter Lang, S. 75-90.

Knoch, Ute/Sitajalabhorn, Woranon (2013): A closer look at integrated writing tasks. Towards a more focused definition for assessment purposes. In: Assessing Writing, 18 (4), S. 300-308.

Kruse, Otto/Ruhmann, Gabriele (2003): Aus alt mach neu: Vom Lesen zum Schreiben wissenschaftlicher Texte. In: Kruse, Otto/Jakobs, Eva-

Maria/Ruhmann, Gabriele (Hrsg.): Schlüsselkompetenz Schreiben. Konzepte, Methoden, Projekte für Schreibberatung und Schreibdidaktik an der Hochschule. Bielefeld: Universitätsverlag Webler, S. 109-121.

Lee, Yong-Won (2006): Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. In: Language Testing, 23 (2), S. 131-166.

Plakans, Lia (2013): Assessment of integrated skills. In: Chapelle, Carol (Hrsg.): The Encyclopedia of Applied Linguistics. Malden, MA: Wiley-Blackwell, S. 205-212.

Stezano Cotelo, Kristin (2008): Verarbeitung wissenschaftlichen Wissens in Seminararbeiten ausländischer Studierende. Eine empirische Sprachanalyse. München: Iudicium.

Weir, Cyril J. (2005): Language testing and validation: An evidence-based approach. Basingstoke, UK: Palgrave McMillan.

Yu, Guoxing (2013): The use of summarization tasks. Some lexical and conceptual analyses. In: Language Assessment Quarterly, 10 (1), S. 96-109.

Kontaktdaten der Autorinnen

Sonja Zimmermann
TestDaF-Institut
Universitätsstr. 134
44799 Bochum
sonja.zimmermann@testdaf.de

Daniela Marks
TestDaF-Institut
Universitätsstr. 134
44799 Bochum
daniela.marks@testdaf.de

Anikó Brandt, Astrid Buschmann-Göbels,
Claudia Harsch (Hrsg.)

Der Gemeinsame Europäische Referenzrahmen für Sprachen und seine Adaption im Hochschulkontext

6. Bremer Symposion zum Sprachenlernen und -lehren

Bibliografische Informationen Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliografische Daten sind im Internet unter *http://dnb.ddb.de* abrufbar.

Printed in Germany
ISBN 978-3-925453-66-3