Håkansson Ramberg, Maria: **Validität und schriftliche Sprachkompetenz. Eine Studie zur Bewertung schriftlicher Leistungen im Fach Deutsch an schwedischen Schulen.** Berlin: Peter Lang, 2023. – ISBN 978-3-631-87372-4. 334 Seiten, € 64,95 [Open Access unter www.peterlang.com].

Besprochen von **Karl-Walter Florin:** Waltrop

https://doi.org/10.1515/infodaf-2024-0026

Die Bewertung schriftlicher Leistungen gehört zu den wesentlichen Aufgaben von Sprachlehrkräften im Schulwesen. Dabei übernehmen diese Lehrkräfte eine hohe Verantwortung, da sie es in den meisten Fällen allein sind, die diese Leistungen feststellen. In ihrer Dissertation *Validität und schriftliche Sprachkompetenz*, erstellt an der Universität Uppsala im Fach Deutsch, untersucht Håkansson Ramberg, wie schwedische Lehrkräfte schriftliche Leistungen ihrer Lernenden bewerten und vergleicht diese Bewertungen mit denen externer Bewerter(innen). Ihr Ziel ist es, die durch die schwedischen Lehrkräfte zugrunde gelegten sprachlichen, textuellen und aufgabenbezogenen Bewertungskriterien zu ermitteln und deren Gebrauch zu vergleichen. Es wird dargestellt, inwieweit die schwedischen Bewertenden übereinstimmen und wie die Bewertungen sich zu den Niveaus des Gemeinsamen europäischen Referenzrahmens (GER) verhalten.

Ausgangspunkt für Håkansson Rambergs Überlegungen ist die Tatsache, dass der GER für den Sprachunterricht eine immer größere Rolle spielt. Die im Referenzrahmen entwickelten Prinzipien und Kriterien nehmen immer stärker Einfluss auf den Fremdsprachenunterricht und bestimmen Bildungspläne sowie Bildungsstandards für die modernen Fremdsprachen in den europäischen Staaten mit. Deshalb ist die Darstellung des schwedischen Bildungswesens und vor allem die Stellung der modernen Fremdsprachen darin wichtige Voraussetzung für die Einordnung der schriftlichen Fremdsprachenkompetenz.

Das Fach Deutsch kann in der Regel ab der 6. Grundschulklasse (von neun) als zweite Fremdsprache gewählt werden, Englisch ist die erste Fremdsprache. Deutsch hat an Bedeutung verloren und ist nach Spanisch die zweithäufig gewählte zweite Fremdsprache. Der Unterricht erfolgt auf sieben Niveaustufen; allerdings kommen die beiden höchsten Stufen sehr selten vor, und selbst die Stufe 5 wird nur selten belegt (39). Das Erreichen bestimmter Stufen ist abhängig vom Beginn des Erlernens der zweiten Fremdsprache. Schwedische Lernende können ab der 6. Klasse, ab der 8. Klasse oder mit Übertritt zum Gymnasium (ab der 10. Klasse) diese in ihren Stundenplan aufnehmen. In der Grundschule sind für das Erreichen einer Stufe zwei Jahre vorgesehen, am Gymnasium pro Stufe ein Jahr. Die schwedischen Lehrpläne beschreiben inhaltliche Bereiche, die behandelt werden sollen,

und sind ergebnisorientiert. Die Lehrpläne legen für die einzelnen Stufen Mindestanforderungen fest, die von den Lernenden zum Ende der Stufe erreicht sein sollen. Håkansson Ramberg stellt schwerpunktmäßig die zentralen Inhalte und Mindestkriterien „hinsichtlich Produktion und Interaktion in den schwedischen Bildungsstandards für Tyska 3, Tyska 4 und Tyska 5“ (45/47) dar. Das Erreichen der Mindestanforderungen gewährleistet das Bestehen des jeweiligen Kurses auf der jeweiligen Stufe, solange die Note E (auf einer Skala von A bis F) vergeben wird. Die Anforderungen in den zweiten Fremdsprachen sind ähnlich dem GER kompetenz- und handlungsorientiert formuliert, allerdings sind die Kriterien allgemeiner gefasst.

Die Bewertung der Leistungen für die Abschlusszeugnisse erfolgt am Ende der Grundschule (9. Klasse) und am Ende der 12. Klasse des Gymnasiums durch die unterrichtenden Lehrkräfte. Diese können auf zentral zur Verfügung gestelltes Material zugreifen (es gibt aber keine zentralen Abschlussprüfungen) und sind weitgehend allein für die Bewertung der erreichten Leistung zuständig. Schwedische Lehrkräfte nutzen für die Feststellung der schriftlichen Leistung eher holistische Bewertungsverfahren und sie können sich dabei des zentral verfügbaren Materials bedienen. Zwar dient dieses dazu, die Bewertungen zu vereinheitlichen, allerdings lässt das dezentrale System durchaus Zweifel an der Gerechtigkeit der Benotung aufkommen.

Für die Autorin stellt sich die Frage, in welcher Beziehung das schwedische Stufensystem zu den Niveaustufen des Gemeinsamen europäischen Referenzrahmen stehen. Nach ausführlicher Darstellung der Grundlagen, Prinzipien und Ziele des GER und der daraus abgeleiteten Testverfahren (Testentwicklung, Testdurchführung und Testbewertung) stellt sie für Schweden fest, dass hier bislang keine zuverlässigen Aussagen über die Vergleichbarkeit der Bewertungen vorliegen. Dabei gibt es inzwischen verschiedene standardisierte Tests für Deutsch (z.B. Deutsches Sprachdiplom, TestDaF oder Goethe Zertifikate). Die erreichten Stufen nach den schwedischen Bildungsstandards werden dennoch zu den GER-Niveaustufen in Beziehung gesetzt, so dass für die Stufen Tyska 3, 4 und 5 die GER-Niveaus A2.2, B1.1 und B1.2 angenommen werden (vgl. Tabelle 6, 67).

Für die Untersuchung der schriftlichen Leistungen im Bereich der zweiten Fremdsprache im schwedischen Schulsystem bezieht sich Håkansson Ramberg auf die Kompetenzmodelle, die dem GER und den schwedischen Bildungsstandards zugrunde liegen. Damit schafft sie die Basis, um Kriterien festlegen zu können, mit denen die Bewertungen der schriftlichen Leistungen durch die Lehrkräfte nachvollzogen und verglichen werden können. Mit Hilfe unterschiedlicher Validitätskonzepte soll die Zuverlässigkeit der Bewertungen überprüft werden. Dabei geht es nicht nur darum zu zeigen, dass ein valider Sprachtest das misst, was er vorgibt zu

messen. Validiert werden soll der gesamte Prozess von der Testkonstruktion über die Durchführung bis hin zur Bewertung und die daraus folgenden Konsequenzen. Dazu greift die Autorin auf zwei Validierungsmodelle zurück: zum einen auf die argumentbasierten Ansätze nach Kane und zum anderen auf das sozialkognitive Rahmenmodell nach Weir, die ausführlich dargestellt werden.

Neben der Validität ist die Reliabilität eines Tests ein weiteres Qualitätsmerkmal für dessen Beurteilung. Dabei kann sowohl die Stabilität der Leistung eines Lernenden als auch die Übereinstimmung von Bewertenden betrachtet werden. Für Håkansson Ramberg geht es vor allem um die Übereinstimmung von Bewertungen durch verschiedene Bewertende. Hier spielen Faktoren eine Rolle wie die Gewichtung von Bewertungskriterien, die Zentraltendenz (also Bewertungen, die zur Mitte tendieren) oder Strenge-/Milde-Effekte. Insgesamt kommt die Autorin zu dem Ergebnis, dass sie bei ihrer Validitätsanalyse einen „sogen. Mixed-Methods-Ansatz“ (102) aus quantitativen und qualitativen Methoden verwenden wird, ohne Vollständigkeit bei den einzelnen Aspekten zu beanspruchen.

Wie dieser Ansatz aussieht, beschreibt die Autorin ausführlich im Kapitel „Forschungsdesign und Forschungsmethodik“ (ab 125) Danach plant sie vier Schritte: Ausgehend von der Datenerhebung führt sie eine qualitative und eine quantitative Analyse der Daten durch, um anschließend die Ergebnisse zu vergleichen und zu interpretieren (Ablaufschema, 126). Durch dieses Vorgehen möchte sie unterschiedliche Validitätsaspekte erfassen und darstellen. Die Daten wurden an schwedischen Gymnasien erhoben, an denen Lernende Deutsch auf den Fremdsprachenstufen Tyska 3, 4 und 5 nach den schwedischen Lehrplänen belegen konnten. Als Vergleichsbasis wird der erfolgreiche Abschluss auf Stufe Tyska 5 in Beziehung zum Niveau B1.2 nach GER gesetzt. Als Aufgabe für die Prüfung wählt Håkansson Ramberg das „Modul Schreiben zur Prüfung Goethe-Zertifikat auf B1-Niveau“ (132), deren erfolgreiche Bearbeitung als B1-Niveau definiert ist. Die Prüfung besteht aus drei Teilaufgaben, die in ihren Anforderungen klar formuliert sind und sich von den schwedischen Aufgaben unterscheiden.

Die Bewertenden setzen sich aus drei Gruppen zusammen: Zum einen gehören schwedische Deutschlehrkräfte dazu, die die Lernenden an den beteiligten Gymnasien unterrichteten (insgesamt 18 Personen). Zum anderen wurden zwei schwedische Bewertende ausgewählt, die als ausgebildete und erfahrene Gymnasiallehrkräfte die Tests als Externe beurteilen. Schließlich wurden zwei zertifizierte muttersprachliche GER-Bewertende mit unterschiedlicher Korrekturerfahrung im Bereich B1 beauftragt. Die schwedischen Lehrkräfte benutzten die für die modernen Fremdsprachen an schwedischen Schulen gebräuchlichen Bewertungsfaktoren (Inhalt, Sprache, Ausdrucksfähigkeit), während die GER-Bewertenden das Bewertungsraster für die Goethe-Zertifikatsprüfung B1 mit den vier Dimensionen Erfüllung, Kohärenz, Wortschatz und Strukturen anwendeten.

Für die Analyse der schriftlichen Leistungen lagen insgesamt 225 Texte vor, die nach ihrer Bewertung einem Auswahlverfahren unterzogen wurden, so dass schließlich jeweils 20 Texte aus den Stufen Tyska 3, 4 und 5 ausgewählt wurden; zugleich wurden für jede Stufe zu gleichen Teilen Texte mit den Bewertungen F, E, C und A bestimmt.

Um auf die drei Forschungsfragen – 1. Bewertungskriterien der drei Bewertergruppen, 2. Bewerterübereinstimmung, 3. Beziehung zwischen schwedischen Fremdsprachenstufen und B1-Niveau nach GER – Antworten zu bekommen, wurden die Kommentare der schwedischen Bewertenden in zehn Hauptkategorien (mit etlichen Subkategorien) kodiert, um sie statistisch auswerten zu können. Darüber hinaus wurden die verschiedenen qualitativen und statistischen Verfahren kritisch erläutert und ihre Grenzen bestimmt.

Die nächsten drei Kapitel widmen sich ausführlich der Darstellung der drei Forschungsfragen. Bei der Analyse der Bewertungskriterien ist festzustellen, dass schwedische Lehrkräfte im Vergleich zu den GER-Bewertenden mehr und unterschiedliche Kategorien für die Bewertung nutzen. Insgesamt werden aber von den einzelnen Lehrkräften weniger und allgemeinere Kategorien verwendet (vgl. Tab. 164). Dies wird in der zusammenfassenden Tabelle (198) deutlich: Die schwedischen Bewertenden verwenden die folgenden Bewertungsdimensionen: formale Strukturen, Wortschatz, Angemessenheit, pauschale Beurteilung Sprache, Aufgabenerfüllung, Verständlichkeit und Gesamteindruck, während die GER-Bewertenden sich auf fünf Dimensionen beschränken: Angemessenheit, Wortschatz, formale Strukturen, Aufgabenerfüllung und Verständlichkeit (jeweils in absteigender Reihenfolge).

Bei der Analyse der Bewerterübereinstimmung zwischen den schwedischen Lehrkräften und den beiden externen schwedischen Bewertenden gibt es erhebliche Abweichungen. Die „prozentuale Übereinstimmung“ (PÜ, 155) zwischen den Lehrkräften und der ersten bzw. zweiten externen Bewertenden beträgt lediglich 38 Prozent bzw. 37 Prozent. Die beiden externen Bewertenden stimmen immerhin zu 60 Prozent überein. Statistisch scheinen die Konsens- und Konsistenzwerte nicht zufriedenstellend zu sein. Die qualitative Analyse verweist auch darauf, dass selbst, wenn gleiche Bewertungskriterien verwendet werden, die Gewichtung unterschiedlich sein und es somit zu Abweichungen von bis zu drei Notenstufen kommen kann. Zudem haben die Lehrkräfte, die in der Regel ihre eigenen Lernenden beurteilen, eine Tendenz zur Milde, während die externen Bewertenden sehr viel öfter Leistungen mit F (unterhalb des Mindeststandards) bewerten.

Im dritten Teil ihrer Analyse setzt Håkansson Ramberg die schwedische Bewertung in Beziehung zum B1-Niveau. Es zeigt sich, dass Texte, die von den schwedischen Bewertenden auf der Stufe Tyska 5 mit mindestens der Note E bewertet wurden, in der Regel auch das B1.2-Niveau nach GER erfüllten. Leistun-

gen, die teilweise von den schwedischen Bewertenden mit der Note F bzw. E beurteilt wurden, erreichten nach den GER-Bewertenden ebenfalls das B1-Niveau. Auf den beiden Stufen Tyska 3 und 4 schafften allerdings auch ein Teil die Anforderungen für das B1.2-Niveau. Hier zeigte sich aber, dass die Mindestanforderungen (Note E auf der entsprechenden Stufe) nicht ausreichten, sondern lediglich gute bzw. sehr gute Leistungen (Tyska 3) und befriedigende bis sehr gute Leistungen (Tyska 4) nach den schwedischen Bewertungen erreicht sein mussten.

Nach der ausführlichen Darstellung der Ergebnisse der quantitativen und qualitativen Analyse geht Håkansson Ramberg an die Interpretation. Da ihr Fokus vor allem auf den Bewertenden liegt, diskutiert sie im ersten Schritt die unterschiedliche Handhabung der Bewertungskriterien durch die verschiedenen Bewertergruppen. Unterschiede sieht sie z.B. darin, dass die schwedischen Lehrkräfte eine andere Lehr- und Bewertungstradition haben und tendenziell zu einem holistischen Bewertungsverfahren neigen. Dabei berücksichtigen sie sehr unterschiedliche Kriterien, die zudem individuell gewichtet werden. Die GER-Bewertenden hingegen beziehen sich überwiegend auf das für die Goethe-Zertifikats-Prüfung vorgegebene Bewertungsraster, nutzen dadurch weniger Kriterien, diese dann aber intensiver. Ein Grund für diese einheitlichere Bewertung kann auch an der systematischen Fortbildung und Zertifizierung der GER-Bewertenden liegen, die es so für die schwedischen Lehrkräfte nicht gibt.

Bei der Betrachtung der Bewerterübereinstimmung stellt die Autorin große Differenzen fest. Besonders die Übereinstimmung bei den Noten ist wenig zufriedenstellend und zeigt sich vor allem bei den höheren Benotungen. Es besteht also nur ein geringer Konsens bei der Beurteilung der einzelnen Leistungen. Etwas besser ist allerdings die Konsistenz der Bewertungen; hier wird deutlich, dass die schwedischen Lehrkräfte ein Gefühl dafür haben, in welcher Stelle einer Rangordnung eine Leistung einzuordnen ist. Unabhängig von der mangelnden Bewerterübereinstimmung zeigt der Vergleich der Bewertung der Stufe Tyska 5 und der Bewertung nach den GER-Kriterien, dass die schwedischen Mindeststandards das behauptete Niveau B1.2 nach GER durchaus erreichen. Dies ist durchaus wichtig, wenn es um die Vergleichbarkeit von Kompetenzniveaus geht.

Die Validitätsstudie zur Bewertung schriftlicher Leistungen im Fach Deutsch an schwedischen Gymnasien zeigt, wie aufwendig das Forschungsdesign für eine solche Arbeit ist. Håkansson Ramberg beschreibt ausführlich die Rahmenbedingungen, mögliche Forschungsansätze und Erklärungsmodelle und erläutert die eingeschränkte Aussagekraft und die vielfältigen Einflussfaktoren, die bei der Interpretation der Ergebnisse zu berücksichtigen sind. Für Forschende, die ähnliche Projekte im Blick haben, lohnt auf jeden Fall die Lektüre, zumal die Autorin auf fast 30 Seiten wesentliche aktuelle Literatur zusammengestellt hat. Für die Lesenden, die vor allem an den Ergebnissen interessiert sind, wäre eine auf den

wesentlichen Inhalt reduzierte Kurzfassung wünschenswert, die sie am Ende auf 15 Seiten für die schwedischen Leser(innen) anbietet.