Zur Beurteilung von Schreibleistungen aus Deutsch als Erstsprache in *High-Stakes Tests*. Die Stabilität von Skalendeskriptoren im Bewertungsraster für die österreichische Matura.

Günther Sigott
Hermann Cesnik

Abstract

The present study describes a first step towards validating the rating scale for assessing L1 German writing in the context of the Austrian Matura exam. After describing the process of scale development in the context of the exam reform, it reports on an empirical study into the stability of scale descriptors. The 70 scale descriptors were assessed in terms of their difficulty by a panel of 100 experienced teachers who had not undergone training in the use of the scale. This data served as the basis for studying overall rater agreement, the correspondence of the sequence of empirically scaled descriptors to the intended sequence, and for studying rater agreement on individual descriptors. It was found that using the scale without previous rater training is not recommendable and rater training is indispensable. The highest level on the scale was found to be the most consensual among the assessors. There is relatively high agreement with regard to what constitutes excellence in L1 German writing. The descriptors on the critical pass level were found to function relatively well although at least two descriptors turned out to be unstable and should be focused on in rater training. Overall, a high number of stable descriptors was found, which is remarkable given that the assessors had not yet received training in using the scale. Suggestions for areas of focus in assessor training or minor improvements of the scale are made.

1. Zur Beurteilung von Schreibkompetenz mittels Skalen

Skalen oder Beurteilungsraster werden sowohl im Zweitsprachenbereich als auch im Erstsprachenbereich zur Beurteilung von Schreibleistungen verwendet. Charakteristisch für die Verwendung von Skalen ist es, dass keine diskreten Elemente, wie etwa Fehler gezählt werden, sondern vielmehr versucht wird, den Gesamteindruck, der in der beurteilenden Person erweckt wird, einer Niveaustufe auf einer mehrstufigen Skala zuzuordnen. Skalen sind somit, anders als Tests, die aus einzelnen Testfragen bestehen, Instrumente zur Beurteilung im Sinne eines qualitativen Beschreibungsansatzes. Dies bedeutet jedoch nicht, dass Qualitätsansprüche, die üblicherweise an Beurteilungsverfahren gestellt werden, bei der Verwendung von Skalen nicht relevant wären. Insbesondere wenn Skalen im Rahmen von Qualifikationsprüfungen Verwendung finden, deren Ergebnis wichtige Folgen für die Kandidaten[1] mit sich bringen, ist es notwendig, Klarheit über die Reliabilität und Validität der Beurteilungen zu schaffen.

Anders als in Testverfahren, die auf diskreten Testfragen oder Items beruhen, für deren Beurteilung in der Regel explizite und exakt anwendbare Beurteilungsschlüssel vorliegen, spielt bei der direkten Beurteilung von sprachlichen Leistungen mittels Skalen die Interpretation der Skalen und der darin enthaltenen Formulierungen eine wesentliche Rolle. Die beurteilenden Personen, die fortan als Assessoren bezeichnet werden, können sich oft erheblich in ihrer Interpretation der jeweiligen Skala unterscheiden. Daher spielt die Varianz, die durch die Interpretation der Skala durch die Assessoren in den Beurteilungsprozess einfließt, eine wesentliche Rolle, wie aus verschiedenen Modellen zum Beurteilungsprozess mittels Skalen klar wird. Dies gilt im Prinzip für die Beurteilung sowohl von mündlichen als auch schriftlichen Leistungen in gleicher Weise (z.B. McNamara 1996, Skehan 1998, Bachman 2002, Fulcher 2003).

Neben der Zuverlässigkeit der Assessoren an sich kommt dabei insbesondere den Interpretationen der Niveaubeschreibungen durch die Assessoren besondere Bedeutung zu. Wenn

[1] Zur besseren Lesbarkeit werden durchgehend männliche Formen verwendet. Es sind jedoch immer sowohl weibliche als auch männliche Personen gemeint.

Niveaubeschreibungen von verschiedenen Assessoren unterschiedlich interpretiert werden, entstehen Zufallseffekte bei der Beurteilung, die die Fehlervarianz erhöhen und daher der Reliabilität und in weiterer Folge der Validität der Beurteilung abträglich sind. Deshalb ist es notwendig, herauszufinden, inwieweit die Interpretation einzelner Niveaubeschreibungen über verschiedene Assessoren hinweg schwankt. Das Ausmaß dieser Schwankung kann als Stabilität der Niveaubeschreibungen bezeichnet werden. Ein geringes Maß an Schwankung wird somit hohe Stabilität abbilden, während ein hohes Maß an Schwankung niedrige Stabilität der Niveaubeschreibungen bedeutet. Der vorliegende Artikel beschreibt eine Untersuchung, in der die Stabilität von Elementen von Niveaubeschreibungen einer Skala, die im Rahmen der Reform der österreichischen Matura entwickelt wurde, im Zentrum des Interesses steht. Der Reformprozess, in dessen Rahmen die Skala entwickelt wurde, fand in den Jahren von 2009 bis 2013 statt (Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens (BIFIE) 2013).

1.1. Skalen zur Beurteilung von Texten in der Erstsprache

Skalen zur direkten Beurteilung von Schreibleistungen fanden ab 1970 sowohl für Englisch als Erstsprache als auch für Englisch als Zweitsprache zunehmend Verwendung (Knoch 2009:18). Einer der ersten Ansätze zur systematischen Entwicklung einer Skala für erstsprachliche englische Texte von Studienanfängern in den USA ist Diederich et al. (1961). Eine Faktorenanalyse von Beurteilungen von 300 Schreibperformanzen durch 53 erfahrene Assessoren ergab fünf Dimensionen, auf die die Assessoren Wert zu legen schienen. Die Dimensionen waren *Ideas, Forms, Flavour, Mechanics* und *Wording*. Analytische Skalen zur Beurteilung von Schreibleistungen in elf verschiedenen Erstsprachen, darunter auch Deutsch, von Teilnehmern aus 14 Ländern in der Sekundarstufe wurden später auch von der *International Association for the Evaluation of Educational Achievement (IEA)* (Gorman et al. 1988) entwickelt. Diese Skalen bestanden aus sieben Kerndimensionen, deren englische Bezeichnungen folgendermaßen lauten: *Quality and scope of content; Organization and presentation of content; Style and tone; Lexical and grammatical features; Spelling and orthographic conventions; Handwriting and neatness; Response of the assessor* (S. 49). Sasaki und Hirose (1999) beschreiben die Entwicklung einer analytischen Skala für expositorische Texte in Japanisch als Erstsprache für Studierende an japanischen Universitäten. In der englischen Dokumentation der Skalenentwicklung werden die sechs Dimensionen folgendermaßen benannt: *Clarity of the theme; Appeal to readers; Expression; Organization; Knowledge of language forms; Social* awareness (Sasaki & Hirose 1999, 463).

Allerdings ist im deutschsprachigen Raum die Verwendung von Skalen zur direkten Beurteilung von sprachlichen Leistungen in der Erstsprache Deutsch weniger verbreitet. Zumindest gibt es kaum publizierte Skalen zur Beurteilung von Schreibleistungen in Deutsch als Erstsprache, insbesondere nicht für Schreibleistungen von jungen Erwachsenen am Ende der Sekundarstufe. Im deutschen Sprachraum dürfte sich die Entwicklung und Verwendung derartiger Instrumente, sofern sie überhaupt stattfand, eher im informellen und nicht öffentlich dokumentierten Bereich vollzogen haben. Eine Skala zur Beschreibung von erstsprachlichen Kompetenzen in Deutsch, die ähnlich weite Verbreitung wie der für Zweit- und Fremdsprachen intendierte Gemeinsame Europäische Referenzrahmen für Sprachen (GERS) hat, existiert nicht. Zumindest in Österreich fand die Entwicklung einer Skala zur Beurteilung von schriftlichen Maturaarbeiten, die auf nationaler Ebene Anwendung findet, erst vor einigen Jahren statt (Glaboniat & Sigott 2012). Diese Skala bildet die Grundlage für den Beurteilungsraster, der derzeit zur Beurteilung von Schreibleistungen in der Matura in Deutsch, aber auch in Slowenisch, Ungarisch und Kroatisch als Unterrichtssprache im gesamten Bundesgebiet verwendet wird (bifie 2014).

1.2. Konstruktionsprinzipien für Skalen

Grundsätzlich ist bei der Entwicklung von Skalen zwischen holistischen und analytischen Skalen zu unterscheiden. Holistische Skalen beschreiben Charakteristika von Leistungen umfassend und detailliert auf verschiedenen Niveaus einer einzigen Skala, weshalb bei holistischen Skalen die Unterscheidung zwischen Skala und Dimension nicht notwendig ist. Analytische Skalen hingegen umfassen mehr als eine Dimension. Auf jeder Dimension werden Charakteristika von für die einzelnen Niveaus typischen Leistungen beschrieben. Diese Beschreibungen werden üblicherweise als Deskriptoren bezeichnet.

Für die Entwicklung beider Arten von Skalen können prinzipiell drei Ansätze unterschieden werden, die als intuitiv, theoriegeleitet und empirisch bezeichnet werden können (Berger 2015:65ff.; Fulcher 2010:208-215). Der *intuitive* Ansatz steht für eine Vorgangsweise, die sich hauptsächlich auf die Intuitionen von erfahrenen Lehrenden und Experten stützt. Dabei werden keine empirischen Daten erhoben oder Schreibperformanzen herangezogen. Vielmehr werden im Zuge von Diskussionen in einer Gruppe, die mit der Entwicklung der Skala betraut wurde, Deskriptoren formuliert. Der Prozess ist in der Regel zyklisch. Die ursprünglichen Formulierungen werden oftmals in mehreren Zyklen überdacht und gegebenenfalls überarbeitet. Der *theoriegeleitete* Ansatz hingegen baut auf einem Kompetenzmodell auf, das zum Ziel hat, die Kompetenzen, die zur Erbringung der geforderten Leistungen notwendig sind, zu beschreiben. Dabei wird der Versuch unternommen, die Dimensionen der Skala, sofern eine analytische Skala anvisiert wird, von den Komponenten des Kompetenzmodells abzuleiten. Sofern das Kompetenzmodell auch Informationen zur Entwicklung der Kompetenz oder Kompetenzen enthält, werden diese in die Formulierung der Deskriptoren einbezogen. Der *empirische* Ansatz beruht auf der Bewertung von Deskriptoren oder Schreibperformanzen durch Informanten. Eine mögliche Vorgangsweise besteht darin, fachkundigen Personen Deskriptoren oder Teile von Deskriptoren vorzulegen mit der Bitte, diese in eine sinnvolle Rangordnung zu bringen. Diese Rangordnung kann sodann mit der Anordnung der Deskriptoren im vorher erstellten Skalenentwurf verglichen werden und Diskrepanzen korrigiert werden. Eine andere Vorgangsweise besteht darin, den fachkundigen Personen Schreibperformanzen vorzulegen und die Personen zu bitten, anzugeben, inwieweit die einzelnen Deskriptoren die einzelnen Schreibperformanzen treffend beschreiben. Eine weitere Alternative besteht darin, die Performanzen hinsichtlich ihrer Qualität in Gruppen zu ordnen. Danach wird versucht, die Kriterien, die die Grundlage für die Gruppenbildung darstellen, explizit zu machen. In der Regel geht der empirischen Phase eine intuitive und allenfalls theoriegeleitete Phase voran.

2. Die Entwicklung des österreichischen Beurteilungsrasters für Deutsch als Unterrichtssprache

2.1. Ausgangssituation

Die Deskriptoren für die vorliegende Untersuchung stammen aus dem Entwicklungsprozess, der zur Beurteilungsskala für die teilzentrale Matura aus Deutsch als Unterrichtssprache im österreichischen Schulwesen führte. Dieser Prozess begann 2009, als in Österreich die Einführung eines teilzentralen Prüfungswesens für die Unterrichtssprachen (Deutsch, Slowenisch, Kroatisch, Ungarisch), für die lebenden Fremdsprachen (Englisch, Französisch, Italienisch, Spanisch), für die klassischen Sprachen (Latein und Altgriechisch) und Mathematik beschlossen wurde. Gegenstand der vorliegenden Untersuchung ist die Skala für die Beurteilung von Schreibleistungen in Deutsch als Unterrichtssprache. Für die große Mehrheit der Schüler in Österreich ist Deutsch Unterrichtssprache, wobei die überwiegende Mehrheit Deutsch als Erstsprache erlernt hat.

Ausgangspunkt für die Skalenentwicklung war eine Reform der Aufgabenstellung. Diese Reform ist im größeren Kontext des Paradigmenwechsels zur Kompetenzorientierung zu sehen, der sich im

österreichischen Schulwesen ab etwa 2005 vollzog. Vor dem Reformprozess wurde für die Matura aus Deutsch ein Aufsatz verlangt, dem eine vom jeweiligen Unterrichtenden formulierte Aufgabenstellung zu Grunde lag. Die Beurteilung lag im Wesentlichen im Ermessen des Unterrichtenden. Die Reform brachte eine Neustrukturierung der Prüfung mit sich. Die Basis für diese Neustrukturierung bildet ein Kompetenzmodell, das Kompetenzen für das Gesamtfach Deutsch (bzw. Kroatisch, Slowenisch oder Ungarisch) beschreibt (bifie 2012). Für die schriftliche Prüfung werden insbesondere die folgenden Kompetenzbereiche definiert: Lesekompetenz, Schriftliche Kompetenz, Argumentationskompetenz, Interpretationskompetenz sowie Sach-/Fachkompetenz. Darüber hinaus spielen Sprachbewusstsein und Reflexionskompetenz in allen zuvor genannten Kompetenzbereichen eine Rolle. Lesekompetenz befähigt zur „Ermittlung von Informationen und Positionen aus unterschiedlichen Texten und Bildern, jeweils auch die Fähigkeit und Bereitschaft, die Adressatenorientierung und Intention des Textes mitzureflektieren“ (S. 9). Daher beinhaltet die Aufgabenstellung immer einen oder mehrere längere Inputtexte, auf die bei der Bearbeitung der Aufgaben Bezug genommen werden muss. Schriftliche Kompetenz umfasst „neben der Fähigkeit zur Sprach- und Schreibrichtigkeit jeweils die Fähigkeit und Bereitschaft, einen der jeweiligen Aufgabe angemessenen Text unter Verwendung der dem gewählten Thema angemessenen stilistischen und textuellen Sprachmittel zu verfassen, sowie die Fähigkeit, den eigenen Text adressatengerecht zu formulieren.“ (S. 9). Argumentationskompetenz meint „[die] Kompetenz, zu einer gegebenen oder selbst gestellten strittigen Frage von sozialer, politischer und/oder kultureller Relevanz in Auseinandersetzung mit den Positionen anderer eine eigene Position aufzubauen sowie diese durch Thesen, Begründungen und Beispiele abzusichern; weiters die Fähigkeit, nachvollziehbare und kohärente Argumentationslinien mit sprachlich angemessenen Mitteln zu realisieren.“ (S. 9). Bei der Interpretationskompetenz handelt es sich um „Kompetenzen der Erschließung, Deutung, Beurteilung und Bewertung pragmatischer und poetischer Texte, auch Bilder […]“ (S. 10). Sach-/Fachkompetenz schließlich umfasst „Kenntnisse der relevanten Konzepte und Begriffe des jeweiligen Themas, Kenntnisse von Daten und Zusammenhängen“ (S. 10). Bei der Bearbeitung der Schreibaufgaben interagieren diese Kompetenzbereiche in komplexer Weise.

Nach der neuen Regelung müssen die Kandidaten zwei Schreibaufgaben bearbeiten. Die zu produzierenden Texte müssen den textuellen Anforderungen von Textsorten gerecht werden, die in einem Textsortenkatalog, der neun Textsorten umfasst, aufgelistet und definiert sind (Staud & Taubinger 2011). Die dort definierten Textsorten sind Textanalyse, Textinterpretation, Zusammenfassung, offener Brief, Leserbrief, Empfehlung, Kommentar, Erörterung und Meinungsrede. Die Aufgabenstellung erfolgt nach standardisierten Rahmenvorgaben. Jede der beiden Schreibaufgaben erfordert auch die Bezugnahme auf einen oder mehrere sogenannte Inputtexte. Je nach Aufgabenstellung ist auf diese Inputtexte in unterschiedlicher Form Bezug zu nehmen. Die Gesamtwortzahl für beide Texte zusammen beträgt 810 bis 990 Wörter. Die Länge der beiden Texte variiert je nach Aufgabenstellung innerhalb der vorgegebenen Gesamtwortanzahl. Jede Aufgabenstellung enhält auch drei oder vier sogenannte Arbeitsaufträge, denen der zu verfassende Text gerecht werden muss. Die Erfüllung dieser Arbeitsaufträge ist auch Bestandteil der Beurteilung. Weitere Details zur Aufgabenstellung mit Beispielen sind auf der Bifie-Homepage (bifie 2016) zu finden.

2.2. Entwicklung des Beurteilungsrasters

Der Beurteilungsraster wurde in mehreren Sitzungen einer internationalen Arbeitsgruppe im Zeitraum von 2011 bis 2013 entwickelt. Dieser Arbeitsgruppe gehörten erfahrene Lehrende im österreichischen Sekundarschulwesen, an der Universität tätige Deutschdidaktiker aus Deutschland, Österreich und der Schweiz, sowie ebenfalls an der Universität im Bereich Prüfen und Bewerten tätige Fachleute an. Der Entwicklungsprozess orientierte sich am oben beschriebenen

Kompetenzmodell, war somit theoriegeleitet, beinhaltete aber auch intuitive und empirische Facetten. Allerdings war es aus logistischen und zeitökonomischen Gründen nicht möglich, die drei Ansätze rigoros voneinander zu trennen und in linarer Abfolge zu durchlaufen. Eine erste Diskussionsrunde führte zum Entschluss, eine analytische Skala mit den vier Dimensionen *Aufgabenerfüllung aus inhaltlicher Sicht, Aufgabenerfüllung aus textstruktureller Sicht, Aufgabenerfüllung in Bezug auf Stil und Ausdruck sowie Aufgabenerfüllung hinsichtlich normativer Sprachrichtigkeit* zu entwickeln. Diese analytische Skala sollte für die Beurteilung von Schreibleistungen in allen neun Textsorten grundsätzlich anwendbar sein. Weiters wurde beschlossen, auf jeder Dimension vier Niveaustufen zu definieren. Dabei bedeutet Niveaustufe 1 das einfachste Niveau (Mindestandforderung) und Niveaustufe 4 das schwierigste (exzellente Leistung). Auf den Versuch, eine Niveaustufe 0 für das Nichterreichen der Mindestanforderung zu beschreiben, wurde verzichtet, da sich die Beschreibung zwingend durch die Nichterfüllung der Anforderungen für die Niveaustufe 1 ergibt.

Bei der Formulierung von Deskriptoren wurde zunächst theoriegeleitet und intuitiv vorgegangen. Die Mitglieder der Arbeitsgruppe schlugen Formulierungen für jedes der vier Niveaus auf jeder der vier Dimensionen vor. In Diskussionen wurden diese Formulierungen adaptiert bis ein Konsens gefunden werden konnte. Darauf folgte eine empirische Phase, in der geprüft wurde, inwieweit Schreibperformanzen, die aus der gleichzeitig laufenden Pilotierung der Aufgabenstellungen stammten, durch die formulierten Deskriptoren treffend beschrieben werden konnten. Dies führte zu weiteren Anpassungen der Formulierungen. Schließlich wurde auch das Kompetenzmodell, das auf der Grundlage umfangreicher Literaturrecherchen und Fachdiskussionen entwickelt worden war, im Hinblick auf seine Kompatibilität mit dem entstehenden Beurteilungsraster überprüft. Die Arbeitsgruppe kam zum Schluss, dass das Kompetenzmodell und die Richtlinien zur Aufgabenerstellung mit dem Beurteilungsraster kongruent waren. Der Beurteilungsraster, der aus dieser Arbeit hervorging, ist mit geringen Formulierungsunterschieden die Basis für den derzeit in Verwendung befindlichen Beurteilungsraster. Der Algorithmus für die Berechnung einer Gesamtnote über beide Aufgaben und Dimensionen hinweg ist in bifie 2014 beschrieben. Der vorläufige Beurteilungsraster, aus dem die Deskriptoren für die vorliegende Untersuchung stammen, ist in den Tabellen 4 bis 7 zu finden (Glaboniat & Sigott 2012).

3. Empirische Untersuchung

Der vorläufige Beurteilungsraster stellt einen Konsens der Arbeitsgruppe dar, nicht aber einen Konsens jener Personen, die ihn später zur Beurteilung anwenden sollen. Daher entsteht die Frage, inwieweit die Formulierungen der Deskriptoren auch für jenen Personenkreis sinnvoll und konsensfähig sind, der den Beurteilungsraster später als Prüfer in der Matura anwenden soll. Freilich werden die Lehrenden in der Anwendung des Beurteilungsrasters in Fortbildungsseminaren geschult. Der Schulungsaufwand hängt jedoch in hohem Maß davon ab, wieweit die Teilnehmer hinsichtlich der Interpretation der Deskriptoren bereits vor Schulungsbeginn übereinstimmen. Daher ist es von Interesse, herauszufinden, welche Deskriptoren auch ohne Schulung konsensfähig sind. Nicht konsensfähige Deskriptoren erfordern entweder Revisionen des Beurteilungsrasters oder besondere Beachtung in der Assessorenschulung.

3.1. Fragestellung

Das Ziel der empirischen Untersuchung besteht darin, festzustellen, inwieweit ungeschulte Assessoren darin übereinstimmen, welchem von vier Niveaus einzelne Deskriptoren zuzuordnen sind. Dabei ist es notwendig, die Niveaus zu definieren. Anders als in den lebenden Fremdsprachen kann hier als Bezugspunkt nicht ein Niveau des GERS herangezogen werden, weil der GERS nicht für die Beschreibung und Beurteilung erstsprachlicher Kompetenzen konzipiert ist. Es muss daher von dem

Konzept des in der österreichischen Matura erwarteten Mindestniveaus ausgegangen werden. Eine den Niveaubeschreibungen des GERS entsprechende Beschreibung der Kompetenzen für die schriftliche Matura aus Deutsch in Österreich liegt bislang allenfalls in Ansätzen vor. Eine zentrale Rolle spielt in diesem Zusammenhang das Konzept des minimal kompetenten Kandidaten (Cizek & Bunch 2007:83). Im Kontext der vorliegenden Untersuchung ist dies jener Kandidat, der gerade noch ausreichend fähig ist, um die schriftliche Matura in Deutsch als Unterrichtssprache zu bestehen. Dieser Kandidat wird die Anforderungen von Deskriptoren, die dem Niveau 1 zugeordnet werden, erfüllen, nicht jedoch jene auf den Niveaus darüber. Es muss für die vorliegende Untersuchung somit von einem prototypischen Kompetenzprofil des minimal kompetenten Maturanten ausgegangen werden, hinsichtlich dessen unter den beurteilenden Lehrenden mehr oder weniger Konsens besteht. Auf der Basis des Konzepts des minimal kompetenten Maturanten (für die schriftliche Matura aus Deutsch als Unterrichtssprache) kann nun gefragt werden, wie schwierig es nach Einschätzung des einzelnen Assessors für den minimal kompetenten Maturanten sein wird, den Anforderungen der einzelnen Deskriptoren des Beurteilungsrasters gerecht zu werden. Auf dieser Basis können nun in Anlehnung an Berger (2015:133) für die einzelnen Deskriptoren die folgenden Forschungsfragen formuliert werden:

1. Inwieweit stimmen die Assessoren hinsichtlich der Reihung der Deskriptoren überein?

2. Inwieweit entsprechen die aggregierten Einschätzungen der Deskriptoren bezüglich ihrer Schwierigkeit der im Skalenentwurf intendierten Reihung?

3. Welche Deskriptoren sind instabil, weil die Einschätzungen der Assessoren bezüglich der Schwierigkeit nicht ausreichend übereinstimmen?

3.2. Materialien und teilnehmende Personen

Vor der Erstellung des Fragebogens wurden Deskriptoren, die aus mehreren eigenständigen Komponenten bestehen, in Deskriptorenteile zerlegt. Dies war beispielsweise für den Deskriptor für das höchste Niveau auf der Dimension *Aufgabenerfüllung aus textstruktureller Sicht* der Fall. Der Deskriptor *Leicht nachvollziehbare Binnengliederung, zielgerichteter, sicherer Einsatz von passenden Textorganisatoren; kohärent und frei von Gedankensprüngen, zielgerechter Einsatz von metakommunikativen Mitteln* wurde in folgende vier Deskriptorenteile zerlegt:

- leicht nachvollziehbare Binnengliederung
- zielgerichteter, sicherer Einsatz von passenden Textorganisatoren
- kohärent und frei von Gedankensprüngen
- zielgerechter Einsatz von metakommunikativen Mitteln

Der Einfachheit halber wird für diese Deskriptorenteile fortan, wie für die nicht zerlegten Deskriptoren, die Bezeichnung Deskriptor verwendet. Nach der Zerlegung der komplexen Deskriptoren standen 76 Deskriptoren zur Verfügung. Dabei ist zu beachten, dass einige Deskriptoren doppelt vorkommen, weil sie in aufeinanderfolgenden Niveaus der Skala zu finden sind. Die 76 Deskriptoren stellen somit *tokens*, nicht *types* dar. Im Fall von Deskriptoren, die auf mehr als einem Niveau vorkommen, wurde für die Analyse nur ein Deskriptor verwendet. In wenigen Fällen unterschieden sich die Schwierigkeitseinschätzungen von doppelt vorkommenden Deskriptoren geringfügig. In diesen Fällen wurde der kleinere der beiden Werte in die Analyse einbezogen. Wenn doppelt vorkommende Deskriptoren nur einmal aufgelistet werden, reduziert sich die Anzahl der Deskriptoren auf 70. Diese sind in den Tabellen 4 bis 7 ausgegraut dargestellt. Die Deskriptorenliste mit allen 76 Deskriptoren, einschließlich der doppelt vorkommenden, diente als Grundlage für die Erstellung eines Fragebogens. Die Reihenfolge der Deskriptoren im Fragebogen wurde willkürlich

verändert, um den Teilnehmern nicht die im Skalenentwurf vorgegebene Reihenfolge zu suggerieren. Außerdem wurde darauf geachtet, doppelt vorkommende Deskriptoren möglichst weit voneinander entfernt im Fragebogen zu positionieren. Der Fragebogen wurde im Jänner 2012 an österreichische Deutschlehrende in allen neun Bundesländern per E-Mail versandt. Alle Lehrenden hatten Unterrichtserfahrung in der Sekundarstufe 2 bzw. unterrichteten zu dieser Zeit in der Sekundarstufe 2 und waren im Rahmen des Prüfungsreformprozesses als Aufgabenersteller tätig, nicht aber in die Entwicklung des Beurteilungsrasters eingebunden. In dieser Gruppe von Lehrenden waren alle Schultypen repräsentiert. Dieser Personenkreis umfasste rund 150 Lehrende. Die Assessoren trugen ihre Bewertungen in ein dem E-Mail angehängtes Excel-Datenfile ein und retournierten dieses wieder per E-Mail an eine eigens dafür eingerichtete E-Mail-Adresse. Bis Februar 2012 lagen 117 ausgefüllte Fragebögen vor, die die Basis für die vorliegende Studie bilden. Die Fragestellung im Fragebogen ist in Tabelle 1 wiedergegeben.

Fragebogen zu den Deskriptoren der Skala SRDP Deutsch

Bitte stellen Sie sich einen Maturanten / eine Maturantin vor, der / die gerade kompetent genug ist, um die Matura zu bestehen.

Lesen Sie dann die Deskriptoren und geben Sie auf der Skala von 1 bis 8 an, wie schwierig es Ihrer Meinung nach für diesen Maturanten / diese Maturantin ist, dem jeweiligen Deskriptor gerecht zu werden, also einen Text zu verfassen, auf den der Deskriptor zutrifft.

1	2	3	4	5	6	7	8

sehr einfach sehr schwierig

Geben Sie Ihre Einschätzung durch Eintragen einer Zahl von 1 bis 8 an:

	Deskriptor	Schwierigkeit 1 bis 8
1	Text gedanklich und grafisch der Textsorte angemessen klar strukturiert	
2	Variantenreiche und komplexe, der Textsorte angemessene Satzstrukturen	
3	…..	

Tabelle 1: Fragebogen

Wie ersichtlich, war von den Teilnehmern jeder Deskriptor hinsichtlich des Grades seiner Realisierbarkeit durch den minimal kompetenten Maturanten auf einer Skala von 1 bis 8 zu beurteilen. An dieser Stelle wird die Verbindung der Skalenentwicklung mit der Setzung von Kompetenzstandards besonders deutlich. Die Beurteilung der „Schwierigkeit“ der einzelnen Deskriptoren hängt in erster Linie von der Vorstellung der Kompetenz des minimal kompetenten Maturanten ab, die der jeweilige Assessor hat. Die Daten spiegeln daher einerseits die Einschätzung der Schwierigkeit der Deskriptoren an sich, d.h. unabhängig vom Kompetenzniveau des minimal kompetenten Maturanten, und andererseits allfällige Unterschiede in der Vorstellung der Kompetenz des minimal kompetenten Maturanten zwischen den einzelnen Assessoren wieder.

3.3. Resultate

Zur Analyse der Daten werden sowohl klassische korrelationsstatistische Verfahren als auch Verfahren der probabilistischen Testtheorie verwendet. Die korrelationsstatistischen Verfahren werden als bekannt vorausgesetzt. Grundlegende Konzepte des Multifacetten-Raschmodells (MFRM) werden im folgenden Abschnitt resümiert.

3.3.1. Das Multifacetten-Raschmodell

Das MFRM stellt eine wichtige Grundlage für die Datenanalyse in dieser Untersuchung dar. Das MFRM zählt zur Gruppe der probabilistischen Testmodelle. Im Gegensatz zu den klassischen

(deterministischen) Testmodellen erfolgt die Bestimmung von Assessorenstrenge, Deskriptorenschwierigkeit, Probandenfähigkeit und Skalenqualität mit Hilfe probabilistischer Verfahren. Jeder Beurteilungsvorgang ist eine Interaktion zwischen diesen Facetten und das MFRM ermöglicht die simultane Berücksichtigung und Auspartialisierung dieser. Das MFRM stellt eine Erweiterung des dichotomen und polytomen Raschmodells dar und spielt insbesondere bei der Bewertung von produktiven Fertigkeiten (Schreibleistungen, Sprechleistungen) eine Rolle. Hier erfolgt die Bewertung der Probandenleistung nicht über Zählen sondern über Bewertungen auf einer mehrstufigen Skala.

Das wichtigste Ergebnis aus dem MFRM sind die Fair Measures, Maßzahlen für die Assessorenstrenge, die Deskriptorenschwierigkeit und die Personenfähigkeit (falls vorhanden). Diese Maßzahlen berücksichtigen bereits die jeweils anderen Facetten und ermöglichen somit einen fairen Vergleich von Strenge, Schwierigkeit und Fähigkeit. Dieser Vorgang wird als Auspartialisieren bezeichnet. Fair Measures sind also Maßzahlen, die wieder auf die originale Bewertungsskala (hier: 1 bis 8) rücktransformiert sind. Damit sind die Ergebnisse auch für den statistisch weniger versierten Leser klar interpretierbar. Modellintern wird aus rechnerischen Gründen allerdings eine logarithmische Skala (Measure) verwendet.

Der Itemfit ist ein wesentliches Kriterium zur Einschätzung der Güte und somit der Plausibilität der Ergebnisse aus dem Raschmodell. Der Itemfit für den Deskriptor beispielsweise zeigt die Übereinstimmung aller auspartialisierten Assessoren in der Bewertung dieses Deskriptors. Sinngemäß gilt dies auch für alle anderen Facetten. Prinzipiell gilt, je höher die Übereinstimmung, desto höher die Beurteilungsgüte. Unterschieden wird nach Infit und Outfit. Der Outfit (Outlier fit) basiert auf einer traditionellen Chi-Quadrat Statistik (quadrierte Abweichung zwischen Beobachtungs- und Erwartungswert). Das Quadrieren der Abweichungen bewirkt, dass Ausreißer die Maßzahl stärker beeinflussen. Der Outfit ist ausreißersensitiv. Der Infit (Information fit) basiert ebenfalls auf einer Chi-Quadrat Statistik, die jedoch mittels Itemvarianz (=Information) gewichtet wird. Der Infit ist varianzsensitiv. Als statistische Maßeinheit zur Einschätzung des Fits dient der Mean Square (MNSQ). Der Wertebereich des MNSQ liegt zwischen Null und plus Unendlich. Werte von oder um Eins sind optimal, weiter entfernt liegende Werte zeigen Misfit zwischen Modell und Empirie (Realität) auf. MNSQ weit unter Eins sind Hinweise auf Overfit oder Abhängigkeiten zwischen den Facettenelementen. Erwünscht ist voneinander unabhängiges Beurteilungsverhalten der Assessoren. MNSQ weit über Eins sind Indikatoren auf Underfit, also das Vorhandensein zusätzlicher Varianz in den Daten (“Rauschen”). Für die Interpretation des MNSQ gelten nach Linacre (2014) folgende Richtwerte: MNSQ < 0,5 (Item kann verwendet werden); MNSQ 0,5 bis 1,5 (Item ist gut geignet); MNSQ > 1,5 bis 2 (Item ist eventuell zu modifizieren); MNSQ > 2 (Item ist zu verwerfen oder zu überarbeiten).

3.3.2. Plausibilitätsprüfung der Daten

Nach Einlangen der 117 Fragebögen wurden die Daten zunächst auf Plausibilität geprüft. Dazu wurde eine Rasch-Multifacettenanalyse mit FACETS (Linacre 1997) mit den Facetten Assessor und Deskriptor durchgeführt. Assessoren, deren Infit oder Outfit Mean Square über 1,5 lag, wurden als Ausreißer identifiziert. Diese Assessoren zeigen ein atypisches Ratingverhalten, das im Widerspruch zum Verhalten des Großteils der anderen Assessoren steht. Diese 16 Assessoren wurden daher aus dem Datensatz eliminiert. Ebenso ausgeschlossen wurden Assessoren, deren mittlere Korrelation mit allen anderen Assessoren kleiner als 0 war. Dies trifft auf einen Assessor zu. Nach der Plausbilitätsprüfung verblieben 100 Assessoren im Datensatz, der als Grundlage für die weiteren Analysen dient.

3.3.3. Übereinstimmung der Assessoren (Interrater Reliability)

Ein grobes Bild der Assessorenübereinstimmung kann durch die Prüfung der internen Konsistenz der Gruppe der 100 Assessoren gewonnen werden. Dies ist grundsätzlich durch die Berechnung des Cronbach-Alpha-Koeffizienten möglich. Alpha wird in der klassischen Testtheorie routinemäßig zur Prüfung der internen Konsistenz von itembasierten Tests verwendet. In dieser Untersuchung kann Alpha herangezogen werden, um die Konsistenz der Assessoren untereinander zu überprüfen. In der Tat liegt der Wert für Alpha sowohl für den Gesamtpool der Deskriptoren wie auch für die Deskriptoren auf jeder der vier Dimensionen getrennt betrachtet durchgehend über 0,9. Allerdings ist bei der Interpretation von Alpha als Maß für interne Konsistenz Vorsicht geboten, da Alpha in hohem Maß von der Größe der Stichprobe, hier der Assessorenanzahl, abhängt. Daher wurden in Anlehnung an Berger (2015:138ff.) weitere Maßzahlen für die Assessorenübereinstimmung berechnet. Die Schwierigkeitsbeurteilungen durch die 100 Assessoren wurden mittels Kolmogorov-Smirnov-Anpassungstest auf Normalverteilung geprüft. Die Prüfung ergab, dass für den Großteil der Assessoren die Beurteilungen nicht normalverteilt sind. Dies legt für die Prüfung der Assessorenübereinstimmung die Verwendung von nichtparametrischen Verfahren nahe. Daher wurden in Anlehnung an Berger (2015:138ff.) der Spearman-Koeffizient sowie Kendall Tau-b berechnet. Um ein globales Maß für die Assessorenübereinstimmung zu erhalten, wurden die Koeffizienten der jeweiligen Koeffizientenmatrix gemittelt (vgl. Berger 2015:139). Die Mittelwerte der Koeffizienten sind in Tabelle 2 dargestellt.

	Spearman Rho	Kendall Tau-b
Aufgabenerfüllung aus inhaltlicher Sicht (IN)	0,39	0,32
Aufgabenerfüllung aus textstruktureller Sicht (TS)	0,31	0,27
Aufgabenerfüllung in Bezug auf Stil und Ausdruck (ST)	0,46	0,39
Aufgabenerfüllung hinsichtlich normativer Sprachrichtigkeit (NR)	0,28	0,25
Total (n=100)	0,37	0,31

Tabelle 2: Mittelwerte Spearman Rho und Kendall Tau-b.

Tabelle 2 zeigt die Mittelwerte der beiden Korrelationskoeffizienten, berechnet auf der Basis der Deskriptoren getrennt nach den vier Dimensionen *Aufgabenerfüllung aus inhaltlicher Sicht* (IN), *Aufgabenerfüllung aus textstruktureller Sicht* (TS), *Aufgabenerfüllung in Bezug auf Stil und Ausdruck* (ST) und *Aufgabenerfüllung hinsichtlich normativer Sprachrichtigkeit* (NR). Weiters wurden die mittleren Koeffizienten auf der Basis des gesamten Deskriptorensets berechnet. Die ausgewiesenen Korrelationskoeffizienten sind Mittelwerte aus allen 4950 maximal möglichen bivariaten Korrelationskoeffizienten [(100*100-100)/2] bei n=100 Assessoren. Wie ersichtlich, liegen die Mittelwerte der paarweisen Korrelationskoeffizienten für Spearman Rho zwischen 0,28 (NR) und 0.46

(ST). Kendall Tau-b schwankt von 0,25 für (NR) bis 0,39 für ST. Die Koeffizienten für das gesamte Deskriptorenset sind 0,37 für Spearman Rho und 0,31 für Kendall Tau-b.

Die Stärke der Korrelation zwischen Assessoren hängt sowohl von Unterschieden in der Rangordnung ab, in die die einzelnen Assessoren die Deskriptoren bringen, als auch von Unterschieden in der Assessorenstrenge. Die Nichtnutzung der gesamten Bandbreite der Schwierigkeitsskala kann zu Boden-, Mitten- und Deckeneffekten führen, die Auswirkungen auf die Stärke der Korrelation haben können. Vereinzelt können einzelne Assessorenpaare auch eine negative Korrelation aufweisen. Alle diese Effekte beeinflussen die durch Korrelationskoeffizienten ausgedrückte Assessorenübereinstimmung in Tabelle 2.

Global betrachtet kann von einer schwachen bis mäßigen mittleren positiven Korrelation zwischen den Assessoren in jeder der vier Deskriptorkategorien (IN, TS, ST, NR) sowie in den Deskriptoren als Gesamtkategorie gesprochen werden. Betrachtet man die Assessorenübereinstimmung in den vier Dimensionen, so wird klar, dass die Übereinstimmung in der Dimension *Aufgabenerfüllung in Bezug auf Stil und Ausdruck* am höchsten ist. Die mittleren Koeffizienten liegen hier bei 0,46 (Spearman Rho) und 0,39 (Kendall Tau-b). Die zweithöchste Assessorenübereinstimmung zeigt sich für die Dimension *Aufgabenerfüllung aus inhaltlicher Sicht*, wo die Koeffizienten bei 0,39 (Spearman Rho) und 0,32 (Kendall Tau-b) liegen. Die dritthöchste Assessorenübereinstimmung besteht in der Dimension *Aufgabenerfüllung aus textstruktureller Sicht*, wo die Koeffizienten bei 0,31 (Spearman Rho) und 0,27 (Kendall Tau-b) liegen. Die geringste Assessorenübereinstimmung zeigt die Dimension *Aufgabenerfüllung hinsichtlich normativer Sprachrichtigkeit*. Hier liegen die Koeffizienten bei 0,28 (Spearman Rho) und 0,25 (Kendall Tau-b). Es zeigt sich somit, dass die Assessoren das Konstrukt *Stil und Ausdruck* am zuverlässigsten interpretieren. Im Gegensatz dazu besteht bei der Interpretation des Konstrukts *normative Sprachrichtigkeit* die größte Unsicherheit unter den Assessoren.

In Summe kann festgehalten werden, dass die Assessoren in der Interpretation der Konstrukte, die durch die analytische Beurteilungsskala operationalisiert werden, zu einem gewissen Grad übereinstimmen. Der Grad der Übereinstimmung wäre jedoch für die Verwendung der Skala in einer Qualifikationsprüfung wie der Matura, aus deren Ergebnissen für die Kandidaten bedeutende Konsequenzen erwachsen, nicht ausreichend. Um ein ausreichendes Maß an Übereinstimmung zu gewährleisten, müssen die Lehrenden in der Anwendung der Skala entsprechend geschult werden. Das Ergebnis zeigt somit, dass Assessorenschulungen für die Verwendung der Skala im operativen Testbetrieb unumgänglich sind und unterstreicht somit die Wichtigkeit der Durchführung von Schulungen zur Skalenverwendung, wie sie im Rahmen der SRDP Unterrichtssprachen routinemäßig stattfinden. Besondere Beachtung sollte dabei der Dimension *normative Sprachrichtigkeit* geschenkt werden, wo die in dieser Untersuchung ungeschulten Assessoren am wenigsten übereinstimmen.

3.3.4. Übereinstimmung der Deskriptoren mit der intendierten Skala

Die Grundlage für die Prüfung der Übereinstimmung der Deskriptoren mit der intendierten Skala ist die Reihung der Deskriptoren auf der Basis der Schwierigkeitseinschätzungen durch die Assessoren. Da diese Einschätzungen unter anderem von Unterschieden in der Assessorenstrenge abhängen, wurden die Daten einer Rasch-Multifacettenanalyse unterzogen, wodurch Unterschiede in der Assessorenstrenge auspartialisiert werden. Um einen Vergleich der Positionierungen der Deskriptoren auf der Basis der Schwierigkeitseinschätzungen aus der Rasch-Multifacettenanalyse mit der intendierten Positionierung in der Skala zu ermöglichen, wurde die nach Schwierigkeitseinschätzung geordnete Liste der 70 Deskriptoren in vier Sektoren unterteilt, die den vier Niveaus in der Skala entsprechen. Von den 70 Deskriptoren befinden sich jeweils 18 in den

Niveaus 1, 2 und 3 und 16 in Niveau 4. Wenn die Schwierigkeitseinschätzungen vollkommen mit der Positionierung der Deskriptoren in der Skala übereinstimmen, müssen sich alle Deskriptoren aus Niveau 1 in Sektor 1, jene aus Niveau 2 in Sektor 2, jene aus Niveau 3 in Sektor 3 und jene aus Niveau 4 in Sektor 4 wiederfinden. In diesem Fall wären die Sektorengrenzen kongruent mit den Niveaugrenzen in der Skala. Tabelle 3 zeigt die nach Schwierigkeitseinschätzung geordnete Deskriptorenliste sowie die Sektorengrenzen zwischen den vier Niveaus. Auf dieser Grundlage kann man erkennen, welche Deskriptoren auch auf der Basis der Schwierigkeitseinschätzungen dem jeweiligen intendierten Niveau auf der Skala zugeordnet werden. Auf der Basis der Abweichung von dieser Zuordnung können die einzelnen Deskriptoren hinsichtlich ihrer Stabilität beurteilt werden.

Grundlage für die Beurteilung der Deskriptorenstabilität ist die Distanz zwischen der Position, die auf der Beurteilung der Schwierigkeit beruht und der intendierten Position in der Skala. Deskriptoren wurden als instabil betrachtet, wenn die Schwierigkeitsbeurteilung einen Deskriptor in eine Position bringt, die mehr als einen Sektor von der in der Skala intendierten Position liegt. Somit ist beispielsweise ein Deskriptor, der in der Skala auf Niveau 4 liegt, in der nach Schwierigkeitsbeurteilung geordneten Liste jedoch in Sektor 2, als instabil anzusehen.

Weiters wurden auch sogenannte mäßig stabile Deskriptoren identifiziert. Dazu wurden die Sektoren auf der nach Schwierigkeitseinschätzung geordneten Liste halbiert. Deskriptoren, die mehr als einen Halbsektor von ihrer intendierten Position in der Skala entfernt sind, wurden als mäßig stabil betrachtet. Beispielsweise wird damit ein Deskriptor, der in der Skala auf Niveau 2 liegt, in der geordneten Liste jedoch im ersten Sektor von Niveau 1, als mäßig stabil erachtet. In Tabelle 3 sind die instabilen Deskriptoren dunkelgrau und die mäßig stabilen Deskriptoren hellgrau schattiert. Weiters sind in Tabelle 3 auch jene Deskriptoren grau schattiert, für die ungenügende Assessorenübereinstimmung besteht. Details dazu sind unten, Abschnitt 3.3.5. beschrieben.

Total Score	Total Count	Observed Average	Fair Measure	Measure	Infit MNSQ	Outfit MNSQ	Num	kritische Position	Sektor	Deskriptor
217	98	2,21	2,06	-2,1	1,82	1,87	15		Sektor 1 (einfach)	ST1_7u
248	97	2,56	2,43	-1,79	1,24	1,17	3			IN1_3u
284	95	2,99	2,88	-1,46	3,32	3,25	12			ST1_4u
328	100	3,28	3,22	-1,24	1,5	1,44	14			ST1_6u
342	99	3,45	3,43	-1,11	1,01	0,99	7			TS1_2u
331	95	3,48	3,46	-1,09	1,08	1,08	20	int. Niv. 2		IN2_2u
368	100	3,68	3,66	-0,97	0,97	0,97	21	int. Niv. 2		IN2_3u
386	100	3,86	3,85	-0,85	0,81	0,8	6			TS1_1u
378	98	3,86	3,86	-0,84	1,68	1,63	9			ST1_1u
387	99	3,91	3,92	-0,81	1,59	1,57	13			ST1_5u
396	99	4	4	-0,76	0,69	0,68	22			IN2_4u
418	100	4,18	4,2	-0,64	0,77	0,76	5			IN1_5u
415	99	4,19	4,2	-0,64	1,04	1,02	25			TS2_2u
425	100	4,25	4,27	-0,6	1,16	1,15	2			IN1_2u
426	100	4,26	4,28	-0,59	1,1	1,08	4			IN1_4u
438	98	4,47	4,52	-0,45	0,69	0,69	1			IN1_1u
460	100	4,6	4,65	-0,37	0,76	0,73	18			NR1_3u
459	99	4,64	4,68	-0,35	0,53	0,51	40	int. Niv. 3		IN3_4u
463	99	4,68	4,73	-0,33	0,68	0,65	19		Sektor 2	IN2_1u
478	100	4,78	4,85	-0,26	0,94	0,92	33			ST2_6u
474	99	4,79	4,87	-0,24	0,55	0,54	10			ST1_2u
480	99	4,85	4,91	-0,22	0,65	0,66	24			TS2_1u
487	99	4,92	4,99	-0,17	1,41	1,39	23			IN2_5u
472	96	4,92	5	-0,17	0,84	0,83	36			NR2_3u
492	100	4,92	5	-0,17	0,68	0,66	44	int. Niv. 3		TS3_3m

Total Score	Total Count	Observed Average	Fair Measure	Measure	Infit MNSQ	Outfit MNSQ	Num	kritische Position	Sektor	Deskriptor
494	99	4,99	5,07	-0,12	0,54	0,54	30			ST2_3u
502	100	5,02	5,11	-0,1	0,95	0,97	42	int. Niv. 3		TS3_1m
506	100	5,06	5,15	-0,07	0,53	0,51	11	int. Niv. 1		ST1_3u
506	100	5,06	5,15	-0,07	1,16	1,14	39			IN3_3m
491	97	5,06	5,15	-0,07	0,61	0,6	43			TS3_2u
497	98	5,07	5,18	-0,05	0,78	0,79	27			TS2_4u
503	99	5,08	5,19	-0,05	0,64	0,64	26			TS2_3u
506	99	5,11	5,2	-0,04	0,89	0,88	16	int. Niv. 1		NR1_1u
511	100	5,11	5,2	-0,04	0,79	0,78	31			ST2_4u
513	100	5,13	5,23	-0,03	1,51	1,48	38			IN3_2m
514	100	5,14	5,24	-0,02	0,61	0,59	8	int. Niv. 1		TS1_3u
515	100	5,15	5,25	-0,01	0,82	0,79	37		Sektor 3	IN3_1m
519	100	5,19	5,29	0,01	0,93	0,9	58	int. Niv. 4		IN4_4u
522	100	5,22	5,32	0,03	1,08	1,07	34			NR2_1u
520	99	5,25	5,35	0,05	0,55	0,54	28			ST2_1u
514	98	5,24	5,36	0,06	0,75	0,76	32			ST2_5u
531	100	5,31	5,42	0,09	0,76	0,77	47			ST3_1u
534	100	5,34	5,45	0,11	0,88	0,88	59	int. Niv. 4		IN4_5u
538	100	5,38	5,49	0,14	0,63	0,59	62	int. Niv. 4		TS4_2u
542	100	5,42	5,54	0,17	1,09	1,05	54			NR3_3u
534	98	5,45	5,56	0,18	0,94	0,93	64			TS4_4u
534	98	5,45	5,59	0,21	0,64	0,62	49			ST3_3u
544	99	5,49	5,61	0,22	1,35	1,35	35	int. Niv. 2		NR2_2u
547	99	5,53	5,64	0,24	1,28	1,29	17	int. Niv. 1		NR1_2u
549	99	5,55	5,68	0,27	0,62	0,6	51			ST3_5u
556	99	5,62	5,74	0,3	0,92	0,91	73			ST4_7u
570	100	5,7	5,83	0,37	0,78	0,78	69			ST4_3u
566	99	5,72	5,85	0,39	0,95	0,92	45			TS3_4u
573	100	5,73	5,86	0,39	0,6	0,6	29	int. Niv. 2		ST2_2u
576	99	5,82	5,95	0,45	1,2	1,24	52		Sektor 4 (schwierig)	NR3_1u
583	100	5,83	5,96	0,47	0,76	0,71	46			TS3_5m
596	98	6,08	6,22	0,67	1,41	1,33	76			NR4_3u
591	97	6,09	6,23	0,67	0,88	0,85	67			ST4_1u
626	100	6,26	6,4	0,81	1,06	1	41			IN3_5u
624	99	6,3	6,43	0,84	1,46	1,5	53			NR3_2u
632	100	6,32	6,46	0,86	0,91	0,85	50			ST3_4u
640	99	6,46	6,59	0,99	1,02	1,08	48			ST3_2u
653	100	6,53	6,66	1,06	1,33	1,3	75			NR4_2u
642	98	6,55	6,69	1,09	1,02	1	70			ST4_4u
655	99	6,62	6,74	1,14	1,18	1,18	68			ST4_2u
664	100	6,64	6,77	1,17	1,78	1,68	74			NR4_1u
669	100	6,69	6,81	1,22	1,58	1,47	72			ST4_6u
653	97	6,73	6,83	1,24	1,52	1,4	66			TS4_6u
673	100	6,73	6,85	1,26	1,11	1,08	60			IN4_6u
698	98	7,12	7,22	1,75	1,33	1,2	71			ST4_5u

Tabelle 3: Instabile und mäßig stabile Deskriptoren geordnet nach Schwierigkeitseinschätzung.

Total Score: Summe der Rohwerte aller Bewertungen; Total Count: Anzahl der Bewertungen; Observed Average: Mittelwert aller Bewertungen; Fair Measure: auspartialisierter Mittelwert aller Bewertungen auf der Originalskala; Measure: auspartialisierter Mittelwert aller Bewertungen auf der Logit Skala; Infit MNSQ: Maß für Beurteilungsübereinstimmung, varianzfokussiert; Outfit MNSQ: Maß für Beurteilungsübereinstimmung, ausreißerfokussiert; Num: Nummer des Deskriptors im Fragebogen; kritische Position: Items, die außerhalb der definierten Sektorgrenzen liegen; Sektor: Gliederung in vier den Niveaustufen auf der Skala entsprechende Schwierigkeitsstufen; Deskriptor: Kurzbezeichnung des Deskriptors.

Wie aus Tabelle 3 ersichtlich, sind 14 Deskriptoren in Bezug auf die Übereinstimmung mit der intendierten Position näher zu betrachten. Entsprechend den festgelegten Kriterien sind davon zwei Deskriptoren als instabil und 12 Deskriptoren als mäßig stabil anzusehen. Aus der Kurzbezeichnung der Deskriptoren kann die Dimension (IN, TS, ST, NR), das intendierte Niveau in der Skala (1, 2, 3, 4) sowie die Position des Deskriptors in der jeweiligen Dimension (1-7) in der Matrix in den Tabellen 4 bis 7 abgelesen werden. Außerdem ist ersichtlich, ob es sich um einen einfach (u) oder mehrfach (m) vorkommenden Deskriptor handelt. Somit zeigt sich beispielsweise, dass Sektor 1 dreizehn Deskriptoren enthält, die sich auch in der Skala auf Niveau 1 befinden. Fünf Deskriptoren in Sektor 1 befinden sich auf der Skala jedoch auf einem anderen Niveau. Deskriptor IN3_4u befindet sich beispielsweise in der Skala auf Niveau 3. Er wird somit von den Assessoren als wesentlich einfacher zu erfüllen beurteilt als in der Skala intendiert. Ähnlich verhält es sich mit IN2_2u und IN2_3u, die in der Skala auf Niveau 2 stehen, von den Assessoren aber als einfacher zu erfüllen beurteilt wurden. Die anderen beiden Deskriptoren in Sektor 1, die nicht aus Niveau 1 in der Skala stammen, sind IN2_4u und TS2_2u. Sie sind nicht grau schattiert, weil der Grad ihrer Abweichung noch innerhalb der oben definierten Toleranzgrenzen liegt.

In den Tabellen 4 bis 7 sind die Deskriptoren in einer Matrix angeordnet, die der Skala entspricht. Die vier Dimensionen (IN, TS, ST, NR) sind in den Zeilen zu finden. Die Spalten enthalten die vier Niveaustufen. Alle in Tabelle 3 ausgewiesenen instabilen und mäßig stabilen Deskriptoren werden in den Tabelle 4 bis 7 kommentiert. Die Kommentare sind in Kursivschrift eingefügt und hellgrau schattiert. Kommentare, die sich auf die Übereinstimmung der Deskriptoren mit der intendierten Position in der Skala beziehen, sind mit K-Rang gekennzeichnet. Die Kommentare die sich auf die Assessorenübereinstimmung beziehen sind mit K-Fit gekennzeichnet.

Betrachtet man die Kommentare zur Übereinstimmung der Deskriptoren mit der intendierten Rangordnung (K-Rang), so wird klar, dass semantische Überschneidungen von Formulierungen in der Interpretation durch die Assessoren präsent sind. Dies trifft zu auf *„erfüllt“* vs. *„bearbeitet“; „überwiegend“* vs. *„weitgehend“* vs. *„durchgehend“*; *„elementar“* vs. *„zentral“*; *„leicht nachvollziehbar“* vs. *„gut erkennbar“*; *„überwiegend kohärent“* vs. *„gut erkennbare Kohärenz“* vs. *„leicht nachvollziehbar“*. Beachtung verdient auch der Aspekt Lesekompetenz in Verbindung mit dem Inputtext. Hier dürften unter den Assessoren unterschiedliche Auffassungen hinsichtlich der Schwierigkeit von Lesekompetenzen bestehen. Die Assessoren scheinen sich darüber uneinig zu sein, ob das *„Erfassen der Kernaussage“* schwieriger oder einfacher als das *„Erfassen von Einzelaussagen“* ist. Möglicherweise wird *„Kernaussage“* konzeptuell mit dem Erfassen des Gesamtthemas des Textes assoziiert, während das *„Erfassen von Einzelaussagen“* mit Präzision und Detail und damit mit höheren Anforderungen assoziiert wird. Es ist plausibel, dass das Erkennen des Themas eines Textes einfacher ist als das Verstehen von mehreren Detailaussagen. Beachtung verdient auch der Begriff *„nachvollziehbar“*. Dieser könnte mit ‚diffus' und ‚gerade noch erkennbar' assoziert werden und dadurch als leicht zu realisieren betrachtet werden. Schließlich dürften nicht prämodifizierte Formen in manchen Fällen ikonisch interpretiert werden. Prämodifizierte Formen werden unter Umständen auf Grund ihrer größeren Länge als schwieriger betrachtet als nicht prämodifizierte Formen, die kürzer sind. Dies dürfte auf Formulierungen zutreffen wie *„überwiegend der Textsorte angemessen strukturiert“* vs. *„der Textsorte angemessen strukturiert“*; *„weitgehend gelungen“* vs. *„gelungen“* oder *„weitgehend präzise“* vs. *„präzise“*. Bei der Assessorenschulung sollte auf diese Bereiche besonders geachtet werden.

	Niveau			
	1	**2**	**3**	**4**
Aufgabenerfüllung aus inhaltlicher Sicht (IN)	1 (IN1_1u) Schreibhandlung(en) im Sinne der Textsorte überwiegend erkennbar	19 (IN2_1u) Schreibhandlung(en) im Sinne der Textsorte weitgehend realisiert	37 (IN3_1m) Schreibhandlung(en) im Sinne der Textsorte durchgehend realisiert	55 (IN4_1u) Schreibhandlung(en) im Sinne der Textsorte durchgehend realisiert
	2 (IN1_2u) Alle Arbeitsaufträge angesprochen und 3 (IN1_3u) mindestens zwei bearbeitet	20 (IN2_2u) Alle Arbeitsaufträge angesprochen und mindestens zwei erfüllt *K-Rang: „erfüllt" wird als nur geringfügig schwieriger beurteilt als „bearbeitet" und wandert dadurch auf Schwierigkeitsniveau 1.*	38 (IN3_2m) Alle Arbeitsaufträge erfüllt *K-Fit: nur minimal über der definierten Toleranzgrenze; aus diesem Grund nicht kommentiert*	56 (IN4_2u) Alle Arbeitsaufträge erfüllt
	4 (IN1_4u) Wichtige Einzelaussagen/-aspekte des Inputtexts erfasst	21 (IN2_3u) Kernaussage des Inputtexts erfasst *K-Rang: Erfassen der „Kernaussage" wird als weniger schwierig angesehen als Erfassen von „Einzelaussagen" und wandert dadurch auf Schwierigkeitsniveau 1.*	39 (IN3_3m) Inputtext vollständig erfasst	57 (IN4_3u) Inputtext vollständig erfasst
	5 (IN1_5u) In elementaren Punkten überwiegend sachlich richtig	22 (IN2_4u) In elementaren Punkten weitgehend sachlich richtig	40 (IN3_4u) In zentralen Passagen durchgehend sachlich richtig	58 (IN4_4u) Sachlich richtig
			K-Rang: werden um nahezu zwei Niveaustufen einfacher beurteilt als intendiert; dürfte einerseits an der semantischen Ähnlichkeit der Begriffe „elementar" und „zentral" sowie andererseits zwischen der Ähnlichkeit von „überwiegend", „weitgehend" und „durchgehend" liegen.	
		23 (IN2_5u) Ansätze zur Eigenständigkeit	41 (IN3_5u) Über den Inputtext hinaus eigenständig	59 (IN4_5u) Nachvollziehbare Entwicklung eines eigenen Standpunktes *K-Rang: wird als zu einfach beurteilt; „nachvollziehbar" wird vermutlich als abmindernd empfunden*
				60 (IN4_6u) Komplexität und Ideenreichtum

Tabelle 4: Deskriptorenmatrix mit Kommentaren zur Dimension Aufgabenerfüllung aus inhaltlicher Sicht

	Niveau			
	1	**2**	**3**	**4**
Aufgaben-erfüllung aus textstruktureller Sicht (TS)	6 (TS1_1u) gedankliche Grobstruktur des Textes erkennbar	24 (TS2_1u) Text gedanklich und formal überwiegend der Textsorte angemessen strukturiert	42 (TS3_1m) Text gedanklich und formal der Textsorte angemessen klar strukturiert *K-Rang: Wird als zu einfach beurteilt. Die nicht prämodifizierte Form dürfte nur als geringfügig schwieriger beurteilt werden als die prämodifizierte Form („überwiegend“).*	61 (TS4_1u) Text gedanklich und formal der Textsorte angemessen klar strukturiert
	7 (TS1_2u) Erkennbare Bezugnahme auf den Inputtext	25 (TS2_2u) Eindeutige Bezugnahme auf den Inputtext	43 (TS3_2u) Weitgehend gelungene Verknüpfung mit dem Inputtext	62 (TS4_2u) Gelungene Verknüpfung mit dem Inputtext *K-Rang: wird als zu einfach eingeschätzt; möglicherweise wegen fehlender Prämodifikation (ikonische Interpretation)*
	8 (TS1_3u) Überwiegend kohärenter Aufbau innerhalb der Absätze *K-Rang: wird als schwieriger beurteilt; „überwiegend“ dürfte einen höheren Schwierigkeitsgrad ausdrücken als „gut erkennbar“ oder „nachvollziehbar“*	26 (TS2_3u) Gut erkennbare Kohärenz innerhalb der Absätze, 27 (TS2_4u) nachvollziehbarer Einsatz von Kohäsionsmitteln	44 (TS3_3m) Leicht nachvollziehbare Binnengliederung, *K-Rang: wird als zu einfach eingeschätzt; „leicht nachvollziehbar“ und „gut erkennbar“ sind schlecht voneinander zu unterscheiden.* 45 (TS3_4u) zielgerichteter, sicherer Einsatz von Kohäsionsmitteln; 46 (TS3_5m) kohärent und frei von Gedankensprüngen	63 (TS4_3u) Leicht nachvollziehbare Binnengliederung, 64 (TS4_4u) zielgerichteter, sicherer Einsatz von passenden Textorganisatoren; 65 (TS4_5u) kohärent und frei von Gedankensprüngen, 66 (TS4_6u) zielgerechter Einsatz von metakommunikativen Mitteln *K-Fit: Die Assessoren dürften unsicher über die Bedeutung von „metakommunikativen Mitteln“ sein.*

Tabelle 5: Deskriptorenmatrix mit Kommentaren zur Dimension Augabenerfüllung aus textstruktureller Sicht

	Niveau			
	1	**2**	**3**	**4**
Aufgabenerfüllung in Bezug auf Stil und Ausdruck (ST)	9 (ST1_1u) Ansätze zur schreibhandlungs- und situationsadäquaten Sprachverwendung *K-Fit: „Ansätze" dürfte ein weites Spektrum abdecken, daher große Streuung unter den Bewertungen.*	28 (ST2_1u) Überwiegend schreibhandlungs- und situationsadäquate Sprachverwendung	47 (ST3_1u) Weitgehend schreibhandlungs- und situationsadäquate Sprachverwendung	67 (ST4_1u) Durchwegs schreibhandlungs- und situationsadäquate Sprachverwendung mit gegebenenfalls entsprechenden Stilmitteln
	10 (ST1_2u) In den Schlüsselbegriffen treffend, 11 (ST1_3u) überwiegend angemessene und semantisch korrekte Ausdrucksweise *K-Rang: Die Prämodifikation „überwiegend " verleitet dazu, den Deskriptor als schwieriger zu beurteilen als die nicht prämodifizierte Form des Deskriptors.* 12 (ST1_4u) geringe Varianz in Wortwahl *K-Fit: Es mag für manche Assessoren unklar sein, ob geringe Varianz ein erstrebenswertes oder nicht erstrebenswertes Textmerkmal darstellt.*	29 (ST2_2u) Weitgehend präzise Wortwahl und *K-Rang: Die Prämodifikation „weitgehend" verleitet dazu, den Deskriptor als schwieriger zu beurteilen als die nicht prämodifizierte Form des Deskriptors.* 30 (ST2_3u) angemessene und semantisch korrekte Ausdrucksweise, 31 (ST2_4u) erkennbare Varianz in Wortwahl	48 (ST3_2u) Präzise und variantenreiche Wortwahl, 49 (ST3_3u) idiomatisch, dem Inhalt und der Textsorte entsprechend	68 (ST4_2u) Durchgehend differenzierte und variantenreiche Wortwahl, dem Inhalt und der Textsorte entsprechend; 69 (ST4_3u) Verwendung einer angemessenen Fachsprache; 70 (ST4_4u) idiomatisch gewandt; 71 (ST4_5u) feinere Bedeutungsnuancen auch bei komplexeren Sachverhalten deutlich
	13 (ST1_5u) In Ansätzen erkennbare Varianz in Satzstruktur *K-Fit: „Ansätze" dürfte ein weites Spektrum abdecken, daher große Streuung unter den Bewertungen.*	32 (ST2_5u) Erkennbare Varianz in Satzstruktur	50 (ST3_4u) Weitgehend variantenreiche und komplexe Satzstrukturen	72 (ST4_6u) Variantenreiche und komplexe bzw. der Textsorte angemessene Satzstrukturen *K-Fit: Es dürfte Unklarheit über das Verhältnis von Satzstrukturen und Textsorten bestehen.*
	14 (ST1_6u) An den Inputtext angelehnte Formulierungen, 15 (ST1_7u) vieles wortwörtlich übernommen *K-Fit: „vieles" ist relativ, wird daher unterschiedlich interpretiert*	33 (ST2_6u) Ansätze zu eigenständigen Formulierungen in Bezug auf den Inputtext	51 (ST3_5u) Weitgehend eigenständige Formulierungen in Bezug auf den Inputtext	73 (ST4_7u) Eigenständige Formulierungen in Bezug auf den Inputtext

Tabelle 6: Deskriptorenmatrix mit Kommentaren zur Dimension Augabenerfüllung in Bezug auf Stil und Ausdruck

	Niveau			
	1	**2**	**3**	**4**
Aufgabenerfüllung hinsichtlich normativer Sprachrichtigkeit (NR)	16 (NR1_1u) Überwiegend richtige Anwendung der Prinzipien der deutschen Schreibung *K-Rang: Deskriptor wird schwieriger als intendiert beurteilt. Zwischen „überwiegend“ und „weitgehend“ dürfte nicht unterschieden werden..*	34 (NR2_1u) Weitgehend richtige Anwendung der Prinzipien der deutschen Schreibung	52 (NR3_1u) Richtige Anwendung der Prinzipien der deutschen Schreibung	74 (NR4_1u) In Orthographie nahezu fehlerfrei *K-Fit: „nahezu“ dürfte ein weites Spektrum abdecken, daher große Streuung unter den Bewertungen.*
	17 (NR1_2u) Überwiegend richtige Anwendung der Regeln der Zeichensetzung	35 (NR2_2u) Weitgehend richtige Anwendung der Regeln der Zeichensetzung	53 (NR3_2u) Richtige Anwendung der Regeln der Zeichensetzung	75 (NR4_2u) In Bezug auf Zeichensetzung nahezu fehlerfrei
	K-Rang: Zwischen „überwiegend“ und „weitgehend“ dürfte nicht unterschieden werden. Außerdem werden die Begriffe von den Assessoren als wesentlich schwieriger interpretiert als intendiert.			
	18 (NR1_3u) Überwiegend grammatikalisch korrekt	36 (NR2_3u) weitgehend grammatikalisch korrekt	54 (NR3_3u) Frei von Verstößen gegen mehrere Grammatikregeln	76 (NR4_3u) Grammatikalisch nahezu fehlerfrei

Tabelle 7: Deskriptorenmatrix mit Kommentaren zur Dimension Aufgabenerfüllung hinsichtlich normativer Sprachrichtigkeit

3.3.5. Deskriptorenstabilität nach Assessorenübereinstimmung

Die Stabilität eines Deskriptors hängt auch vom Grad der Übereinstimmung der Schwierigkeitsbeurteilungen der Assessoren für den jeweiligen Deskriptor ab. In der Rasch-Multifacettenanalyse kommt der Grad der Übereinstimmung in den Fitstatistiken der Deskriptoren zum Ausdruck. Die Fitstatistiken (Infit, Outfit) geben an, inwieweit die Assessoren auch nach Auspartialisierung der Assessorenstrenge sich hinsichtlich der Beurteilung eines Deskriptors uneinig sind. Tabelle 3 zeigt die Deskriptoren mit auffälligen Fitstatistiken. Es werden auch hier unterschiedliche Grauschattierungen für instabile Deskriptoren und mäßig stabile Deskriptoren verwendet. Wie ersichtlich, weisen acht Deskriptoren auffällige Fit-Werte auf. Ein Deskriptor (ST1_4u) muss auf dieser Basis als instabil betrachtet werden, während die übrigen sieben Deskriptoren als mäßig stabil zu bezeichnen sind.

In den Tabellen 4 bis 7 sind die Deskriptoren, die hinsichtlich der Assessorenübereinstimmung auffällig sind, ebenfalls kommentiert. Die Kommentare, die sich auf die Assessorenübereinstimmung beziehen, sind mit K-Fit gekennzeichnet. Die geringe Stabilität von diesen Deskriptoren dürfte einerseits durch konzeptuelle Unschärfen in der Interpretation einzelner Termini durch die Assessoren bedingt sein. Beispiele dafür sind *„Ansätze“*, *„vieles“* und *„nahezu“.* Andererseits dürften fachliche Aspekte zu Unklarheiten führen. So ist zu überlegen, ob alle Assessoren mit dem Begriff *„metakommunikative Mittel“* vertraut sind und diesen in ähnlicher Weise interpretieren. Es ist auch zu überlegen, ob *„geringe Varianz“* nicht von manchen Assessoren als ein positives Textmerkmal interpretiert wird im Sinne von ‚frei von Stil- oder Registerbrüchen'. Schließlich könnte auch Unklarheit darüber bestehen, welche Satzstrukturen charakteristisch für welche Textsorten sind. Zur Verbesserung der Assessorenübereinstimmung sollte in der Assessorenschulung explizit auf diese fachlichen Aspekte eingegangen werden. Außerdem sollte bei der Assessorenschulung auf die Bedeutung der Termini *„Ansätze“*, *„vieles“* und *„nahezu“* eingegangen werden. Bei einer eventuellen zukünftigen Skalenrevision wäre hier eine Neuformulierung zu überlegen.

4. Schlussfolgerungen

Die Untersuchung der Assessorenübereinstimmung zeigt, dass die Schulung der Assessoren zur Erhöhung der Übereinstimmung unumgänglich ist. Eine Verwendung der Skala durch nicht geschulte Assessoren in der Durchführung der Prüfung ist nicht ratsam. Das Ergebnis unterstreicht somit die Wichtigkeit der Durchführung von regelmäßigen Assessorenschulungen, wie dies bereits vom Bifie praktiziert wird. Dabei sollte der Tendenz zu unterschiedlicher Assessorenstrenge in der Dimension *Aufgabenerfüllung hinsichtlich normativer Sprachrichtigkeit* besondere Beachtung geschenkt werden. Zusätzlich kann den Problemen durch Revision des Beurteilunsrasters begegnet werden. Ein derartiger Revisionsprozess ist derzeit im Gang und steht kurz vom dem Abschluss.

Die Untersuchung hat weiters ergeben, dass nach den gesetzten Toleranzmaßstäben knapp 70 Prozent der Deskriptoren als stabil bezeichnet werden können. Dieser Prozentsatz ist umso beachtlicher als die Daten von Assessoren stammen, die für die Anwendung der Skala noch keinerlei Schulung durchlaufen hatten. Sektor 1, der im Idealfall sämtliche für das Bestehen der Matura zur erfüllenden Deskriptoren beinhalten müsste, weist mit 61 Prozent den geringsten Prozentsatz an stabilen Deskriptoren auf. Er enthält zwei als instabil eingestufte Deskriptoren, die in der Assessorenschulung besondere Beachtung verdienen und fünf Deskriptoren die mäßig stabil sind, weil sie entweder einem höheren Niveau in der intendierten Skala entstammen und daher als zu einfach beurteilt wurden oder weil die Assessorenübereinstimmung in dem jeweiligen Deskriptor gering ist. Sektor 2 enthält sechs mäßig stabile Deskriptoren. Sektor 3 enthält neben einem instabilen Deskriptor, der besondere Aufmerksamkeit in der Assessorenschulung verdient, fünf mäßig stabile Deskriptoren. Der Prozentsatz an stabilen Deskriptoren ist in beiden Sektoren mit 67 Prozent

identisch. Mit 81 Prozent weist Sektor 4 den höchsten Anteil von stabilen Deskriptoren auf. Es scheint ein hohes Maß an Konsens über die erwarteten Charakteristika von exzellenter Leistung zu geben.

In weiterer Folge wäre es interessant, zu untersuchen, wie gut sich die Deskriptoren zur Beurteilung von authentischen Performanzen eignen. Dazu müssten authentische Performanzen in das Untersuchungsdesign einbezogen werden. Eine derartige Studie würde eine sinnvolle Weiterentwicklung der vorliegenden Untersuchung darstellen.

Die Untersuchung der Beurteilungen der Deskriptoren durch die Assessoren hat es ermöglicht, Aussagen über die Stabilität einzelner Deskriptoren zu tätigen. Damit steht eine empirisch fundierte Basis für Schwerpunktsetzungen in den Assessorenschulungen zur Verfügung.

Bibliographie

Bachman, L. 2002. Some reflections on task-based language performance assessment. *Language Testing* 19(4), 453 – 476.

Berger, A. 2015. *Validating Analytic Rating Scales. A Multi-Method Approach to Scaling Descriptors for Academic Speaking.* Frankfurt am Main: Peter Lang.

bifie 2011. Staud, H. & Taubinger, W. 2011. *Textsortenkatalog.* Klagenfurt/Wien. https://www.bifie.at/node/1498. Abgerufen 20.07.2016.

bifie 2012. Kompetenzmodell. https://www.bifie.at/system/files/dl/srdp_de_positionspapier_2012-04-19.pdf. Abgerufen 20.07.2016.

bifie 2014. Beurteilungsraster SRDP Unterrichtssprache https://www.bifie.at/system/files/dl/srdp_us_beurteilungsraster_2014-11-14.pdf. Abgerufen 20.07.2016.

bifie 2016. Aufgabenbeispiele SRDP Unterrichtssprache. https://www.bifie.at/downloads?projekt[0]=69&schulfach[0]=75&&&dokumenttyp[0]=20&&page=2. Abgerufen 20.07.2016.

Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens (BIFIE) (Hrsg.) 2013. *Standardisierte kompetenzorientierte Reifeprüfung / Reife- und Diplomprüfung. Grundlagen – Entwicklung – Implementierung.* Wien.

Cizek, G.J. & Bunch, M.B. 2007. *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks, CA: Sage Publications.

Diederich, P.B., French, J.W. & Carlton, S.T. 1961. Factors in judgments of writing ability, Research Bulletin 61 – 15. Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction No. ED 002 172.)

Fulcher, G. 2003. *Testing Second Language Speaking.* London: Pearson Longman.

Fulcher, G. 2010. *Practical Language Testing.* London: Routledge.

Glaboniat, M. & Sigott, G. 2012. Das Konzept der kriterienorientierten Bewertung. Dimensionen und Niveaus. *Zeitschrift für den Deutschunterricht in Wissenschaft und Schule* 1, 130 – 140.

Gorman, T.P., Purves, A.C. & Degenhart, R.E., (Hrsgg.) 1988. *The IEA study of written composition I: the international writing tasks and scoring scales.* Oxford: Pergamon Press.

Knoch, U. 2009. *Diagnostic Writing Assessment. The Development and Validation of a Rating Scale.* Frankfurt am Main: Peter Lang.

Linacre, J. M. 2014. *Facets Computer Program for Many-Facet Rasch Measurement, Version 3.71.4.* Beaverton, Oregon: Winsteps.com.

McNamara, T. 1996. *Measuring Second Language Performance.* London: Longman.

Sasaki, M. & Hirose, K. 1999. Development of an analytic rating scale for Japanese L1 writing. *Language Testing* 16 (4), 457 – 478.

Skehan, P. 1998. *A Cognitive Approach to Language Learning.* Oxford: Oxford University Press.

Günther Sigott

Institut für Anglistik und Amerikanistik

Alpen-Adria-Universität Klagenfurt

www.aau.at/~gsigott

Hermann Cesnik

Zentraler Informatikdienst

Alpen-Adria-Universität Klagenfurt

www.aau.at/~hcesnik