Le|se|räu|me
ZEITSCHRIFT FÜR LITERALITÄT
IN SCHULE UND FORSCHUNG

NADJA LINDAUER / TIM SOMMER

Verfahren der Textbeurteilung. Merkmale und Vorzüge eines holistischen Benchmarkratings

Abstract

Die Frage nach adäquaten Verfahren zur Textbeurteilung ist sowohl im Schul- als auch im Forschungskontext von zentraler Bedeutung. Im vorliegenden Beitrag wird ihr vor dem Hintergrund zweier Dissertationsprojekte, in denen es die Qualität von Schülertexten zu erfassen galt, nachgegangen. Mit Blick auf das Erkenntnisinteresse, die verfügbaren Ressourcen und die forschungsrelevanten Gütekriterien werden die Vor- und Nachteile verschiedener Verfahren zur Bestimmung von Textqualität beleuchtet und anschliessend das in beiden Projekten gewählte Verfahren näher vorgestellt. Dabei handelt es sich um das holistische Benchmarkrating, welches im deutschen Sprachraum bisher kaum Beachtung gefunden und sich bezüglich Ressourcen und Gütekriterien als vorteilhaft erwiesen hat.

The question of adequate text quality assessment is of prime importance in school and research. In the present paper, this question will be addressed in the light of two dissertation projects in which the quality of student texts had to be assessed. In view of the research interest, the available resources and the research-relevant quality criteria, the advantages and disadvantages of various methods to assess text quality are elaborated. The method of choice in both projects, the holistic benchmark rating, is presented in more detail. This rating procedure has only received limited attention in the German-speaking world but has proven to be advantageous in terms of resources and quality criteria.

0 | Einleitung

Die Frage nach der adäquaten Erfassung von Textqualität erweist sich in verschiedenen Zusammenhängen als höchst relevant. So ist sie zum einen für Lehrkräfte bedeutsam, welche z. B. wissen möchten, ob ihre SchülerInnen gute Texte verfassen können. Zum anderen kommt ihr aus bildungspolitischer Perspektive bei der Ausarbeitung von Kompetenzstufen und der Überprüfung von Bildungsstandards im Rahmen nationaler Schulleistungsstudien eine gewichtige Rolle zu (vgl. Bundesinstitut bifie 2018; IQB 2016; Konsortium HarmoS Schulsprache 2010; National Center for Education Statistics 2012). Schließlich ist sie auch für Forschende entscheidend, welche beispielsweise ermitteln wollen, ob die von ihnen durchgeführte Intervention im Schreibunterricht wirksam war. Im vorliegenden Artikel wird die Frage nach der Bestimmung von Textqualität aus letztgenannter Perspektive beleuchtet und diskutiert.

In der Forschung finden vor allem zwei Verfahren zur Bestimmung von Textqualität Anwendung: zum einen die holistische Beurteilung, bei der ein Globalurteil abgegeben wird, zum anderen die analytische Kodierung, bei der für verschiedene Kriterien eine einzelne Einschätzung vorgenommen wird (vgl. Neumann 2017, 209 ff.). Bei der Entscheidung für das eine oder andere Verfahren stellt sich die Frage, welcher Ansatz sich vor dem Hintergrund des spezifischen Erkenntnisinteresses, der verfügbaren Ressourcen und der forschungsrelevanten Gütekriterien anbietet.

Diese Frage wird im vorliegenden Beitrag auf Basis zweier Dissertationsprojekte diskutiert, in denen die Erfassung der Qualität von Schülertexten einen zentralen Aspekt darstellt: Projekt A untersucht den Schreibprozess von schriftschwachen Jugendlichen beim Verfassen persuasiver Texte. Zur Ermittlung dieser Jugendlichen wird in einer vorausgehenden Untersuchung bei einer größeren Stichprobe mit einer persuasiven Schreibaufgabe die Textqualität erhoben und auf dieser Grundlage – sowie zusätzlich einer Lehrpersoneneinschätzung – ein selektives Sampling für die Hauptuntersuchung vorgenommen. Die für die zweite Untersuchung ausgewählten schriftschwachen Jugendlichen bearbeiten eine weitere persuasive Aufgabe, wobei zur Erfassung der beim Schreiben ablaufenden mentalen und beobachtbaren Prozesse auf die Methoden des Lauten Denkens, der Videobeobachtung sowie des Smartpens zurückgegriffen wird.

Projekt B ergründet das Schreibwissen von Primarschülerinnen und -schülern und klärt, inwiefern sich das schreibbezogene metakognitive Wissen in verschiedenen Textgenres unterscheidet und ob es mit dem erfolgreichen Bewältigen von entsprechenden Schreibaufgaben zusammenhängt. Für die Beantwortung dieser Fragestellung verfassen SchülerInnen der sechsten Jahrgangsstufe jeweils zwei narrative und zwei instruktive Texte. Im Anschluss daran beantworten sie für beide Genres schriftlich Fragen, die auf das Schreibwissen zielen.

Den beiden Projekten liegen folglich verschiedene Fragestellungen zugrunde, sie haben aber eine ähnliche Ausgangslage, was die forschungsmethodischen Ansprüche und die Ressourcen betrifft. Zur Erfassung der Textqualität fiel die Wahl in beiden Arbeiten auf ein holistisches Rating mit Benchmarktexten, das im deutschsprachigen Raum bisher kaum Beachtung gefunden hat. Dieses Verfahren soll – im Anschluss an einen Überblick über die existierenden holistischen und analytischen Ansätze zur Textbeurteilung (Kapitel 1) sowie eine Diskussion ihrer Vor- und Nachteile (Kapitel 2) – in diesem Beitrag vorgestellt und zudem seine hohe Reliabilität aufgezeigt werden (Kapitel 3 und 4).

1 | Holistische und analytische Verfahren der Textbeurteilung

Wenn es um die Beurteilung von Textqualität geht, wird grundsätzlich zwischen analytischen und holistischen Verfahren unterschieden. Den holistischen Verfahren liegt die Leitfrage zugrunde, was ein guter Text ist. Bei den analytischen Verfahren geht es dagegen eher darum, welche Aspekte einen guten Text ausmachen. Beide Verfahren haben unterschiedliche Ausgestaltungen, auf die im Folgenden eingegangen wird.

Holistische Textbeurteilung: Im Rahmen einer holistischen Textbeurteilung geben die RaterInnen ein Globalurteil ab, das den Gesamttext in den Blick nimmt. Dabei kommt in der Regel eine mehrstufige Skala zum Einsatz, die durch unterschiedliche Kriterien der Textqualität definiert wird. Um ein möglichst hohes Maß an Objektivität zu erlangen, werden der Ratingskala oft beispielhafte Texte beigelegt, welche die unterschiedlichen Qualitätsstufen exemplarisch abbilden.

Ein prominentes Beispiel für ein holistisches Rating ist die im amerikanischen *National Assessment of Educational Progress* verwendete Skala (National Assessment Governing Board 2010, 35 ff.). Auf einer sechsstufigen, textgenrespezifischen Skala werden die Texte mithilfe von unterschiedlichen Kriterien wie z. B. *development of ideas* oder *language facility and conventions* global eingeschätzt. Diese Kriterien sind der primäre Referenzpunkt für die Qua-

litätsbeurteilung. Zusätzlich stehen Beispieltexte zur Verfügung, welche die unterschiedlichen Stufen der Skala repräsentieren und damit die Einordnung erleichtern.

Bei einer anderen Form von holistischen Ratings bilden nicht die beschreibenden Kriterien, sondern Benchmarktexte einen ersten Bezugspunkt. Dabei wird ein Text als Benchmark ausgewählt, der nach vordefinierten Merkmalen als durchschnittlich einzustufen ist. Alle zu beurteilenden Texte werden in der Folge mit diesem Benchmarktext verglichen und eingeordnet. Alternativ können auch drei oder fünf Benchmarktexte herbeigezogen werden, welche unterschiedliche Qualitätsstufen abbilden und so die Beurteilung der Texte vereinfachen. Ein entsprechendes Verfahren haben z. B. Bouwer/Koster et al. (2016) und Feenstra (2014) verwendet. Wenn im Folgenden von Benchmarkrating die Rede ist, wird auf dieses Verfahren, in welchem Benchmarktexte die primären Anhaltspunkte zur Einstufung der Textqualität bilden, Bezug genommen.

Analytische Textbeurteilung: Bei einem analytischen Vorgehen wird nicht der Gesamttext fokussiert. Vielmehr stehen einzelne Dimensionen der Textqualität, wie z. B. Textstruktur, Inhalt oder Sprache, im Fokus, die durch unterschiedliche Items abgebildet werden. Durch die Zusammenführung der dichotom oder mehrstufig eingeschätzten Items liegt am Ende je ein Wert für die verschiedenen Dimensionen der Textqualität vor. Das Vorgehen sieht dabei vor, jeweils ein Item bei allen Texten des Korpus einzuschätzen, bevor das nächste Item berücksichtigt wird. Es wird also nicht Text für Text untersucht, sondern Item für Item. Damit soll der holistische Einfluss, welcher bei der Beurteilung eines ganzen Textes entsteht, verringert werden. Da unterschiedliche Dimensionen der Textqualität beachtet werden, entsteht insgesamt ein differenzierteres Bild der Textqualität als bei einem holistischen Zugang.

Im deutschsprachigen Raum ist das Zürcher Textanalyseraster ein weitverbreitetes und bekanntes Beispiel eines analytischen Instruments (vgl. Nussbaumer 1991). Dieses Raster erlaubt die Beurteilung der Textqualität mithilfe der Dimensionen „Korrelate/Bezugsgrößen", „sprachsystematische und orthografische Richtigkeit" sowie „Angemessenheit", wobei die unterschiedlichen Dimensionen in weitere Unterbereiche unterteilt sind. Diese Unterbereiche werden schließlich durch eine Liste von Items abgebildet. Die Dimension „Angemessenheit" beinhaltet beispielsweise den Unterbereich „Aufbau, Gliederung", welcher weiter differenziert ist in die Items „innere Gliederung" und „äußere Gliederung". Für genre- und aufgabenspezifische Analysen können einzelne Items aus dem Raster ausgewählt, angepasst und unterschiedlich gewichtet werden.

Genre- und aufgabenspezifische Items wurden beispielsweise von Neumann (2007) zur analytischen Kodierung der Schülertexte aus den beiden Schulleistungsstudien DESI (Deutsch-Englisch-Schülerleistungen International) und LAU11/ULME1 (Lernausgangslage Untersuchungen in Hamburg) entwickelt. Die SchülerInnen verfassten unter anderem Reklamationsbriefe zu einer fehlerhaften Computerlieferung, die detailliert mit 47 formalen und inhaltlichen Einzelitems eingeschätzt wurden. Zwei sehr spezifische inhaltliche Items zielten etwa darauf, ob im Rahmen der Beanstandungen – entsprechend der Situationsbeschreibung in der Aufgabenstellung – der fehlende CD-RS-Brenner oder die fehlende Software „Coral Craw" genannt werden.

Die Einteilung in holistische vs. analytische Beurteilungsverfahren wird mitunter nicht einheitlich gehandhabt. Dies zeigt sich insbesondere bei der Erfassung einzelner Textdimensionen auf einer mehrstufigen Skala. Eine entsprechende Vorgehensweise wurde etwa in den Vergleichsarbeiten der 8. Klasse im Kompetenzbereich Schreiben (VERA-8) gewählt (vgl. Schipolowski/Böhme 2016): Neben einem holistischen Gesamteindruck und analytischen Items im vorangehend beschriebenen Sinne wurden die drei Dimensionen „Inhalt", „Stil" und „sprachliche Korrektheit" vierstufig eingeschätzt. Schipolowski und Böhme bezeichnen diese Herangehensweise als semi-holistisch, da der gesamte Text zu berücksichtigen ist, allerdings nur in Bezug auf eine Dimension (vgl. ebd., 8). Im Unterschied zu Schipolowski und Böhme (2016, 8) ordnen z. B. Bouwer/Koster et al. (2016, 3) oder Weigle (2002, 114–119) gleichartige Beurteilungsverfahren den analytischen Verfahren zu.

2 | Entscheidungsgrundlagen bei der Wahl eines Beurteilungsverfahrens

Wenn es im Forschungskontext darum geht, ein geeignetes Verfahren zur Beurteilung der Textqualität zu wählen, müssen vor allem das Erkenntnisinteresse der Studie, die bereitstehenden Ressourcen und die forschungsrelevanten Gütekriterien bei der Entscheidung einbezogen werden. Die folgenden Ausführungen zielen auf diese drei Aspekte.

2.1 | Erkenntnisinteresse

Die Entscheidung für ein Beurteilungsverfahren hängt insbesondere vom Erkenntnisinteresse, welches einem Forschungsprojekt zugrunde liegt, ab (vgl. Neumann 2017, 207). Stehen z. B. die Schreibkompetenzen von Schülerinnen und Schülern im Fokus, kann ein holistischer Blick auf die Textqualität zwar zu der Aussage führen, ob die Kompetenzen hoch oder niedrig sind, allerdings kann nur ein analytisches Rating darüber informieren, welche Teilkompetenzen vertieft erworben sind und welche nicht (vgl. Böhme/Bremerich-Vos et al. 2009). Ein analytischer Zugang bietet folglich einen höheren Informationsgehalt als ein holistischer, da mehrere Dimensionen der Textqualität bestimmt werden. Zugleich ist jedoch umstritten, inwiefern mit den Informationen aus zahlreichen spezifischen Einzelkriterien adäquate Aussagen über die Qualität des Textes als Gesamtes vorgenommen werden können, da vielmehr die einzelnen Items als der globale Text im Fokus stehen (vgl. Kapitel 2.3).

Es stellt sich nun für die eingangs umrissenen Projekte die Frage, welches Beurteilungsverfahren vor dem Hintergrund des Erkenntnisinteresses angezeigt ist. Wie bereits erwähnt, geht es in Projekt A darum, anhand der erhobenen Qualität der Schülertexte – in Kombination mit dem Lehrpersonenurteil – Leistungsgruppen zu bilden und auf dieser Grundlage eine Fallauswahl vorzunehmen: Ermittelt werden sollen schwach schreibende Jugendliche, um deren Schreibprozess in den Blick nehmen zu können. Von Interesse sind nicht Jugendliche mit spezifischen Schwierigkeiten beim Schreiben, zu deren Identifikation einzelne Dimensionen der Textqualität erfasst und damit ein analytisches Verfahren eingesetzt werden müsste. Vielmehr sollen diejenigen Jugendlichen bestimmt werden, deren Texte als Ganzes schwach ausfallen.

In Projekt B geht es um den Zusammenhang von erhobenen Schreibleistungen und dem schreibbezogenen metakognitiven Wissen beim narrativen und instruktiven Schreiben. Das schreibbezogene Wissen soll im Rahmen dieser Studie nicht mit einzelnen Dimensionen der Textqualität, sondern mit den Texten als Ganzes verglichen werden, damit anschließend Rückschlüsse auf den Zusammenhang von Schreibwissen und -leistungen gezogen werden können. Zusammenfassend lässt sich für beide Projekte festhalten, dass die Qualität des Gesamttextes im Zentrum des Interesses steht. In Anbetracht dessen erscheint eine holistische Textbeurteilung als die zielführende Herangehensweise.

2.2 | Ressourcen

Zweitens sind neben dem Erkenntnisinteresse ökonomische Aspekte zu beachten, womit primär zeitbedingte und personelle Ressourcen gemeint sind. Ungeachtet des jeweiligen Verfahrens sind die Planung und die Durchführung von Textbeurteilungen in der Regel mit einem hohen Aufwand verbunden. Hinzukommt, dass mehrere unabhängige Personen die Texte beurteilen müssen. Gerade in Qualifikationsvorhaben kann das eine gewichtige Hürde darstellen, da diese Studien oft alleine durchgeführt werden und weitere Personen zu rekrutieren sind.

Ein zusätzlicher Aufwand ergibt sich durch die Schulung der RaterInnen, in welcher das gewählte Verfahren trainiert wird. In diesem Zusammenhang verweist Canz darauf, dass Schulungen bei einem analytischen Vorgehen häufig kürzer dauern als bei einem holistischen (vgl. Canz 2015, 47). Das mag damit zusammenhängen, dass die einzelnen Kriterien in ei-

nem analytischen Rating in der Regel enger gefasst sind. Des Weiteren gelingt der Zugang zu analytischen Verfahren einfacher, da die Items nacheinander einzeln bewertet werden, oft sogar nur dichotom. In holistischen Ratings hingegen werden unterschiedliche Kriterien für das Globalurteil herbeigezogen und je nach Relevanz unterschiedlich gewichtet. Diese Gewichtung der Kriterien im Rahmen eines Gesamturteils ist komplexer als die Beurteilung von einzelnen Items. Als günstig hat sich in diesem Zusammenhang herausgestellt, wenn eine holistische Beurteilung mithilfe von Benchmarktexten vorgenommen wird. So berichten etwa Bouwer und Kollegen, dass im Fall des von ihnen verwendeten holistischen Benchmarkratings lediglich ein kurzes Training erforderlich war (Bouwer/Béguin et al. 2015, 89).

Während im Hinblick auf die Schulung der RaterInnen folglich keiner der beiden Zugänge als eindeutig überlegen gilt, wird für die anschließende Durchführung der Beurteilung der holistische Zugang übereinstimmend als weniger zeitaufwändig als der analytische beschrieben. In erster Linie ist das damit zu erklären, dass für das holistische Rating ein Globalurteil abzugeben ist. Bei einem analytischen Verfahren müssen hingegen mehrere Einzelkriterien eingeschätzt und die Texte dazu in der Regel mehrfach durchgearbeitet werden, woraus insgesamt eine längere Bearbeitungszeit resultiert (vgl. Weigle 2002, 120).

In einer Qualifikationsarbeit sind die zur Verfügung stehenden Ressourcen oft begrenzt, weshalb sich Verfahren eignen, die nicht zu komplex sind und so weniger Zeit für die Schulung und die eigentliche Textbeurteilung benötigen. Wie bereits angedeutet, wird das holistische Benchmarkrating mit Blick auf Schulung und Durchführung als vergleichsweise ressourcenfreundlich beschrieben (z. B. Schoonen 2005, 10; Bouwer/Béguin et al. 2015, 89) und erscheint deshalb in beiden vorgestellten Projekten als geeignet.

2.3 | Gütekriterien

Drittens gilt es bei der Wahl eines Beurteilungsverfahrens die forschungsrelevanten Gütekriterien zu berücksichtigen. Dazu zählen insbesondere die Objektivität, Reliabilität und Validität, wobei erstere auch häufig als ein Aspekt der Reliabilität verstanden wird (vgl. Grotjahn/Kleppin 2017, 45). Im Folgenden werden daher die beiden Kriterien der Reliabilität und Validität fokussiert.

Die *Reliabilität* stellt als Maß der zuverlässigen Reproduzierbarkeit einer Messung ein zentrales Gütekriterium dar. Im Kontext der Textbeurteilung wird insbesondere die Beurteiler-Reliabilität viel diskutiert. Sie betrifft den Einfluss der beurteilenden Personen auf das Resultat. Textbeurteilungen gelten in diesem Zusammenhang als reliabel, wenn eine Person die Qualität ein und desselben Textes zu verschiedenen Zeitpunkten gleich einschätzt (Intrarater-Reliabilität) oder wenn verschiedene Personen bei ein und demselben Text zum gleichen Urteil gelangen (Interrater-Reliabilität) (vgl. für eine ausführlichere Darstellung und Diskussion des Gütekriteriums der Reliabilität z. B. Grotjahn/Kleppin 2017).

Das Erzielen einer zufriedenstellenden Beurteiler-Reliabilität ist an eine sorgfältige Planung und Durchführung der Textbeurteilung geknüpft und erfordert je nach Vorgehen spezifische Maßnahmen. Diese Maßnahmen untersuchten Graham/Harris et al. (2011, 24) im Rahmen einer Meta-Analyse. Dabei hat sich erstens eine breite Skala (beispielsweise 20 statt 6 Stufen) für die Einschätzung der Texte als zentral herausgestellt. Zweitens haben sich klare Beschreibungen der unterschiedlichen Qualitätsstufen der Skala sowie dazugehörige Textbeispiele als vorteilhaft erwiesen. Die dritte Empfehlung betrifft vor allem holistische Verfahren mit Ankertexten und bezieht sich auf den Beurteilungsablauf: Es hat sich ein zweistufiges Vorgehen bewährt, bei dem der zu beurteilende Text in einem ersten Schritt jenem Benchmarktext zugeordnet wird, dessen Qualität am ähnlichsten erscheint. In einem zweiten Schritt erfolgt eine Präzisierung, indem noch ein Plus oder Minus hinzugefügt bzw. Punkte abgezogen oder hinzugezählt werden (vgl. Graham/Harris et al. 2011, 25).

Dem analytischen Verfahren wird häufig eine höhere Beurteiler-Reliabilität attestiert als dem holistischen (z. B. Bouwer/Koster et al. 2016, 3; Weigle 2002, 120). Um zu beurteilen, welches Verfahren tatsächlich zuverlässiger ist, sind Studien beizuziehen, welche beide Ver-

fahren am selben Textkorpus verwendet haben. Dabei wird im Folgenden auf Untersuchungen zurückgegriffen, welche das holistische Benchmarkrating einbeziehen.

Beispielsweise setzten Van den Bergh/De Maeyer et al. (2012) verschiedene Zugänge zur Beurteilung argumentativer Texte von Studierenden im ersten Jahr der Ausbildung ein. Einerseits unterzogen sie die Texte einem holistischen Benchmarkrating, bei dem alle Texte mit einem Benchmarktext von durchschnittlicher Qualität zu vergleichen und auf einer intervallskalierten Skala mit einem Mittelwert von 100 einzustufen waren. Andererseits nahmen sie eine analytische Kodierung zahlreicher Items zu vier Dimensionen (Struktur, Inhalt, Argumentation, Schlussfolgerung) vor. Die durchschnittliche Übereinstimmung von fünf unabhängigen Personen in Form von Cronbachs Alpha beträgt beim holistischen Vorgehen .82, beim analytischen Vorgehen .88. Der analytische Zugang zeigte folglich eine etwas höhere Interrater-Reliabilität als der holistische.

Zu einem anderen Ergebnis gelangten Bouwer/Koster et al. (2016), welche an zwei persuasiven Aufgaben das holistische Benchmarkrating zum einen mit einem holistischen Verfahren ohne Ankertexte, zum anderen mit einem analytischen Verfahren verglichen. Das holistische Benchmarkrating erfolgte im Unterschied zur Studie von Van den Bergh/De Maeyer et al. (2012) nicht nur mit einem, sondern mit fünf Ankertexten unterschiedlicher Qualität (-2 *SD*, -1 *SD*, *M*, +1 *SD*, +2 *SD*). Beim zweiten holistischen Verfahren war für jeden Text die globale Qualität unter Berücksichtigung der fünf Kriterien „Inhalt", „Struktur", „formaler Aufbau", „Stil" und „sprachformale Aspekte" auf einer Skala von 1 bis 10 zu bestimmen. Im Zuge der analytischen Auswertung wurden 15 Kriterien (je sechs zu Inhalt und Aufbau sowie drei zu Schreibziel bzw. Adressatenorientierung) dichotom eingeschätzt. Die Interrater-Reliabilitäten für drei unabhängig beurteilende Personen lagen beim holistischen Benchmarkrating bei .74 (Aufgabe 1) bzw. .82 (Aufgabe 2), beim kriteriengeleiteten holistischen Rating bei .84 bzw. .74 und beim analytischen Vorgehen bei .84 bzw. .78. Die AutorInnen zeigen mittels Signifikanztests, dass es keine signifikanten Unterschiede zwischen den Verfahren bezüglich der Interrater-Reliabilität gibt. Demnach erwies sich der analytische Zugang in dieser Studie als nicht reliabler als der holistische.

Aufgrund dieser Ausführungen ist festzuhalten, dass bezüglich der Beurteiler-Reliabilität kein Verfahren dem anderen per se vorzuziehen ist. Ferner kann geschlussfolgert werden, dass aufgrund der skizzierten Studien nichts gegen eine Verwendung von holistischen Beurteilungsverfahren mit Benchmarktexten spricht und diese die Textqualität reliabel abbilden können.

Eine zufriedenstellende Reliabilität bildet eine notwendige, aber nicht hinreichende Voraussetzung für die *Validität*, welche als das zentrale Gütekriterium gilt. Unter Validität wird das Ausmaß verstanden, mit dem eine Messung die Kompetenz erfasst, welche sie zu erfassen vorgibt. Geht es um die Wahl eines Verfahrens zur Textbeurteilung, interessiert insbesondere, inwieweit mit einem holistischen und analytischen Zugang dieselbe Kompetenz erfasst wird. Diese Frage betrifft die Konstruktvalidität (vgl. für eine ausführlichere Darstellung und Diskussion des Gütekriteriums der Validität z. B. Canz 2015 oder Grotjahn/Kleppin 2017).

Zur Klärung dieser Frage verglichen Bouwer/Koster et al. (2016) in ihrer bereits erwähnten Studie die drei verwendeten Verfahren (a: holistisches Benchmarkrating, b: holistisches Rating ohne Ankertexte, c: analytische Kodierung) in minderungskorrigierten Korrelationsanalysen miteinander. Dafür summierten die AutorInnen die 15 dichotomen Items des analytischen Instruments zu einem Gesamturteil. Zwischen den beiden globalen Zugängen zeigten sich die höheren Korrelationskoeffizienten (.96 bzw. 1 für die zwei Aufgaben) als zwischen den holistischen und analytischen Einschätzungen (.73–.97). Zusätzlich zu den Korrelationsanalysen haben Bouwer/Koster et al. (2016) Generalisierbarkeitsanalysen durchgeführt, um zu ermitteln, in welchem Maß die für die Textqualität vergebenen Werte die Schreibkompetenz der SchülerInnen oder verschiedene Fehlerarten wie Zufallsfehler oder systematische Fehler, die z. B. in der Aufgabe oder den Raterinnen/Ratern begründet sein können, abbilden (vgl. ausführlicher zu Generalisierbarkeitsanalysen in der Schreibforschung z. B. Schoonen 2012). Die entsprechenden Analysen zeigen, dass im Falle eines von einer Person beurteilten

Textes beim holistischen Rating mit Benchmarktexten 33 %, beim holistischen Rating ohne Ankertexte 23 % und beim analytischen Verfahren 28 % der Varianz in der Textqualität auf die Kompetenz der ProbandInnen zurückgeht. Wenngleich bei allen Verfahren pro SchülerIn mehrere, von verschiedenen Personen beurteilte Texte erforderlich sind, um ein zufriedenstellendes Generalisierbarkeitslevel von .8 oder höher zu erreichen, so manifestiert der mit 33 % höchste Wert beim Benchmarkrating dennoch, dass mit diesen Urteilen besser generalisierbare Aussagen zur Schreibkompetenz getroffen werden können als mit denjenigen aus den anderen zwei Verfahren. Zu ähnlichen Ergebnissen gelangten Schoonen (2005, 2012) sowie Van den Bergh/De Maeyer et al. (2012). Zusammenfassend lässt sich für die Validität somit festhalten, dass der Stand der Forschung auf einen leichten Vorteil des holistischen Ratings mit Ankertexten hinweist. Insofern empfiehlt sich dieses Verfahren für die Beurteilung der Schülertexte in den beiden Dissertationsprojekten.

Vor dem Hintergrund des spezifischen Erkenntnisinteresses, der zeitlichen und finanziellen Ressourcen sowie der Forschungslage zu den Gütekriterien wurde folglich in den beiden Projekten ein holistisches Benchmarkrating gewählt, das in Anlehnung an Bouwer/Koster et al. (2016) und Feenstra (2014) konzipiert ist und diverse evidenzbasierte Empfehlungen aufnimmt (vgl. Graham/Harris et al. 2011). Im nächsten Kapitel werden der Ablauf des Ratings sowie die damit erzielten statistischen Kennwerte berichtet.

3 | Zur Durchführung des holistischen Benchmarkratings

Tabelle 1 gibt einen Überblick über die Stichproben und Schreibaufgaben der beiden Projekte, in welchen das holistische Benchmarkrating zum Einsatz kam.

Projekt	Klassenstufe	Aufgabe/Schreibziel	Textgenre	Textsorte
A (*N*=176)	8 (Niveau Basis- bzw. allgemeine Anforderungen[1])	eine fiktive Firma davon überzeugen, eine unvollständige Sammlung an Punkten im Kontext einer Werbeaktion zu akzeptieren und eine iTunes-Karte zuzuschicken (basierend auf Rijlaarsdam /Braaksma et al. 2008)	Argumentation	Brief
B (*N*=195)	6	zwei vorgegebene Geschichtsanfänge für die MitschülerInnen spannend zu Ende schreiben (basierend auf Schneider/Wiesner et al. 2012)	Narration	Fantasiegeschichte
		mithilfe von Video-/Bilderinput zwei Bastelanleitungen für jüngere SchülerInnen schreiben, damit diese den Gegenstand nur mithilfe der geschriebenen Anleitung basteln können (basierend auf Schneider/Wiesner et al. 2012)	Instruktion	Bastelanleitung

Tab. 1: Überblick über Stichprobe und Schreibaufgaben

Die einzelnen Schritte des Beurteilungsverfahrens sind in Abbildung 1 dargestellt (u. a. basierend auf Feenstra 2014). Ihnen liegt das Ziel zugrunde, ein reliables Rating anhand einer Skala mit fünf Benchmarktexten unterschiedlicher Qualität durchzuführen. Dazu werden die Benchmarktexte im Rahmen verschiedener vorbereitender Schritte empirisch ermittelt. Im

1 Dies entspricht ungefähr der Hauptschule in Deutschland.

Folgenden werden die einzelnen Arbeitsschritte näher dargelegt. Dabei wird exemplarisch auf die persuasive Schreibaufgabe in Projekt A zurückgegriffen. Bei dieser Aufgabe, die in Tabelle 1 bereits umrissen ist und auf Rijlaarsdam/Braaksma et al. (2008) basiert, handelt es sich um eine situierte Aufgabe mit außerschulischem, sozialem Kontext. Ausgangspunkt ist eine laufende Werbeaktion für Schokoriegel der Marke „Schokofreude". Dazu versieht die zuständige Firma jede Verpackung der Schokoriegel mit einem Punkt und stellt bei 20 gesammelten Punkten eine iTunes-Karte im Wert von 50 Franken in Aussicht. Die SchülerInnen sammeln diese Punkte, allerdings sind bereits zehn Tage vor Ende der Aktion keine Schokoriegel mit aufgedruckten Punkten im Handel mehr erhältlich, so dass sie die geforderte Zahl an Punkten nicht erreichen. Aus diesem Grund schreiben die SchülerInnen der Firma „Schokofreude" einen Brief, in welchem sie die Problemlage erklären und die Firma davon zu überzeugen versuchen, dass die unvollständige Punktesammlung nicht ihr Fehler ist und ihnen die iTunes-Karte zusteht.

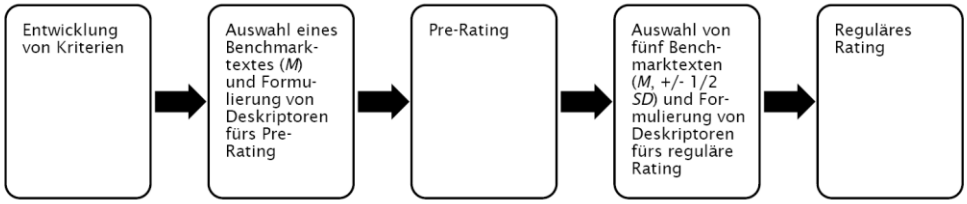

Abb. 1: Arbeitsschritte des holistischen Benchmarkratings

Wird für eine Schreibaufgabe wie die Schokofreude-Aufgabe eine Ratingskala mit Benchmarktexten entwickelt, so gilt es – wie in Abbildung 1 veranschaulicht – vier verschiedene Arbeitsschritte zu durchlaufen, bevor die eigentliche Beurteilung der Textprodukte erfolgen kann. Der *erste Arbeitsschritt* ist die Entwicklung von Kriterien, welche die Textbeurteilung leiten. Sowohl für die Schokofreude-Aufgabe in Projekt A als auch die Aufgaben in Projekt B wurden die drei Dimensionen „Inhalt", „Aufbau" und „Sprache" gewählt und nach den Anforderungen der Aufgaben weiter differenziert. So umfasste die inhaltliche Dimension bei der Schokofreude-Aufgabe zum einen das kommunikative Schreibziel, welches in Form einer Bitte nach der iTunes-Karte, die in der Werbeaktion versprochen wird, vorkommen sollte. Ein anderes Kriterium betraf das inhaltliche Schreibziel. Dabei sollte das Problem verständlich dargelegt, d. h. erklärt werden, dass nicht alle geforderten Punkte geschickt werden können, weil bereits vor dem offiziellen Ende der Aktion keine Punkte mehr im Handel erhältlich sind. Bei der Dimension „Aufbau" wurde u. a. beurteilt, ob der Text überzeugend aufgebaut ist, indem er etwa zunächst in die Situation einführt, danach das Problem entfaltet und mit einer Forderung der iTunes-Karte endet. Schließlich wurde bei der sprachlichen Dimension z. B. in den Blick genommen, inwiefern der Sprachstil adressatengerecht, d. h. höflich und sachlich, ausfällt. Neben der Auswahl der Kriterien wurde in beiden Projekten eine Gewichtung beschlossen: Aufgrund der in den Aufgaben formulierten Schreibziele kommt der inhaltlichen Dimension eine größere Bedeutung zu als den anderen Dimensionen und sie sollte daher bei der Beurteilung stärker gewichtet werden.

Mithilfe der Kriterien und Gewichtung wurde in einem *zweiten Schritt* ein Text von durchschnittlicher Qualität als Benchmarktext für das nachfolgende Pre-Rating ermittelt. Dazu wurden zunächst von einer Person alle Texte nach ihrer Qualität in drei Gruppen („unterdurchschnittlich", „durchschnittlich", „überdurchschnittlich") eingeteilt. In einem zweiten Durchlauf wurden die als durchschnittlich eingeschätzten Texte nach dem gleichen Verfahren nochmals kategorisiert. Von den in beiden Runden als durchschnittlich erachteten Texten wurde anschließend in der Diskussion mit einer zweiten Person ein repräsentativer Text als Benchmark bestimmt. Er erhielt einen Wert von 100 und wurde mit einer Beschreibung seiner Stärken und Schwächen entlang der Kriterien versehen.

In größeren Projekten bietet es sich an, diesen zweiten Schritt im Team vorzunehmen und eine Teilstichprobe aller Texte von mehreren Mitarbeitenden gruppieren zu lassen. Dies ließ sich in den Dissertationen aufgrund der begrenzten finanziellen Ressourcen nicht umset-

zen. Um dennoch sicherstellen zu können, dass ein repräsentativer Text von durchschnittlicher Qualität als Benchmark ausgewählt wird wurde – wie bereits erwähnt – eine Zweitmeinung zu einer kleinen Zahl an Texten, die als Ankertexte in Frage kommen, eingeholt.

Abbildung 2 zeigt das Instrument für das Pre-Rating aus Projekt A. In der oberen Hälfte ist der Benchmarktext aufgeführt, der untere Kasten enthält die Beschreibung.

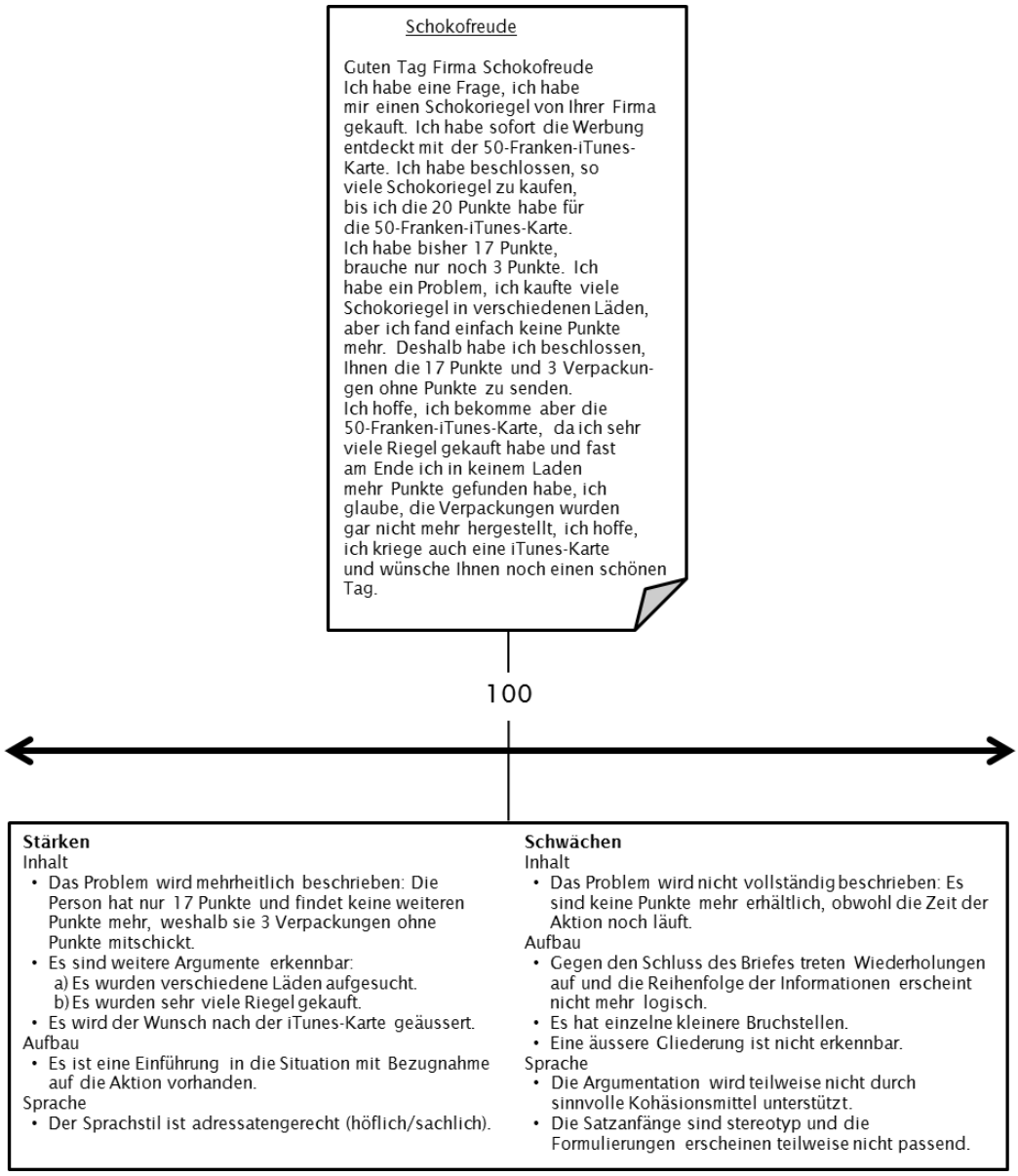

Abb. 2: Benchmarktext und Beschreibung aus dem Pre-Rating in Projekt A

Ins Pre-Rating (*Schritt 3*) waren mehrere unabhängige Personen involviert (Projekt A: 4, Projekt B: 3). Dabei handelte es sich um Lehrkräfte und Forschende mit Erfahrung in der Textbeurteilung. Sie wurden in einem einstündigen Treffen, bei dem mündlich Informationen gegeben und einige wenige Texte bearbeitet wurden, eingeführt. Danach beurteilten sie unabhängig voneinander eine zufällig ausgewählte Teilstichprobe aller Schülertexte (Projekt A: 70, Projekt B: 50). Dabei sollten sie jeden Text mit dem Benchmarktext vergleichen und einschätzen, wie viel besser oder schlechter der zu beurteilende Text als der Referenztext ist. Erschien ein Text z. B. halb so gut wie der Benchmarktext, erhielt er einen Wert von 50. Umgekehrt wurde einer Leistung, die doppelt so gut wirkte, ein Wert von 200 zugewiesen.

Um zu überprüfen, ob die Urteile konsistent ausfallen und damit eine verlässliche Basis für die Auswahl der Benchmarktexte für das reguläre Rating darstellen, wurde die Intraklassenkorrelation (ICC) berechnet. Gemäß Klassifikation von McGraw und Wong (1996) wurde die ICC(C,*k*) Fall 2 (Konsistenz von *k* zufällig ausgewählten Raterinnen/Ratern) herangezogen. Die Interrater-Reliabilitäten fielen mit ICC(C,4) = .84 bei den argumentativen, ICC(C,3) = .80 bei den narrativen und ICC(C,3) = .81 bei den instruktiven Texten zufriedenstellend aus. Um individuelle Beurteilungstendenzen in Form von Milde und Strenge auszugleichen, wurden die vergebenen Werte in einem nächsten Schritt einer Z-Standardisierung an den individuellen Ratermittelwerten unterzogen. Anschließend wurde für jeden Text der Mittelwert kalkuliert. Die Texte wurden diesen Werten gemäß in eine Ranking-Reihenfolge gebracht. In der Folge wurden die Werte transformiert und in eine Skala mit einem Mittelwert von 100 und einer Standardabweichung von 15 Punkten überführt.

Das Ziel dieses Prozesses bildete eine Ratingskala mit fünf Benchmarktexten – einem Text von durchschnittlicher Qualität und vier Texten, deren Qualität eine bzw. zwei Standardabweichungen über und unter dem Durchschnittswert liegt (vgl. Abbildung 3). Demzufolge wurden in einem *vierten Schritt* alle Texte mit Punktzahlen um 70, 85, 100, 115 und 130 herausgefiltert und daraus fünf Produkte ausgewählt, die geringe Abweichungen zwischen den drei Raterurteilen aufwiesen, die jeweilige Qualitätsstufe gut abbildeten und als Ankertexte geeignet erschienen. Für jeden Text wurden die Stärken und Schwächen im Hinblick auf die Dimensionen „Inhalt", „Aufbau" und „Sprache" festgehalten.

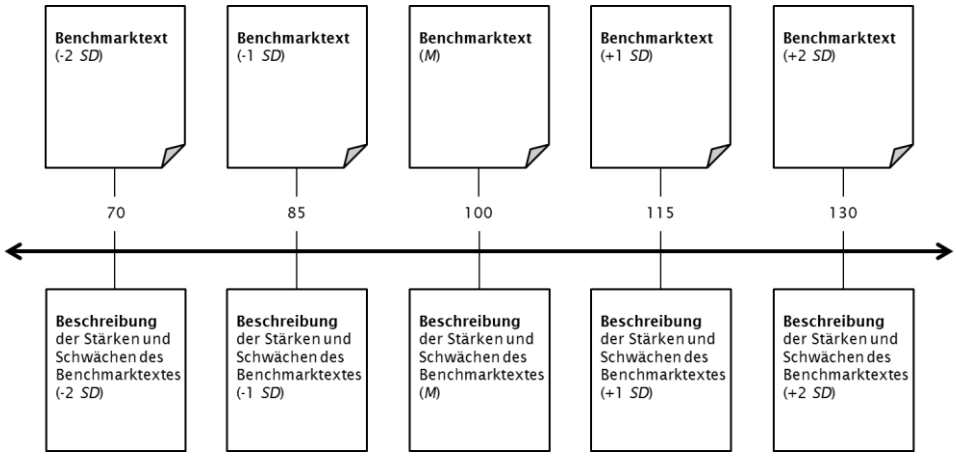

Abb. 3: Aufbau einer Ratingskala mit fünf Benchmarktexten

Der fünfte und letzte Schritt stellte die eigentliche Beurteilung aller Schülertexte dar. Dieser Schritt wurde von drei Personen, die nicht in die vorgängigen Schritte involviert waren, vorgenommen. Einführend fand wiederum ein Training statt, in dem einige Texte (Projekt A: 26, Projekt B: 20) beurteilt und die Einschätzungen besprochen wurden. Aufgabe der RaterInnen war es, jeden Text mit den fünf Benchmarktexten und Beschreibungen zu vergleichen und ihm eine Punktzahl zuzuweisen. Auch Punktzahlen unter derjenigen des schwächsten Benchmarktextes bzw. über derjenigen des stärksten Benchmarktextes waren zulässig. Um zu gewährleisten, dass der Inhaltsdimension stärkeres Gewicht beigemessen wird als den anderen beiden Dimensionen (vgl. Schritt 1), wurde den Raterinnen/Ratern ein gestuftes Vorgehen nahegelegt: Sie sollten zunächst jeden Text aufgrund seines Inhalts einordnen und erst in einem zweiten Schritt Aufbau und Sprache analysieren und ggf. die für den Inhalt vorläufig vergebene Punktzahl korrigieren. Es blieb jedoch den beurteilenden Personen überlassen, um wie viele Punkte sie ihre erste inhaltliche Einschätzung anglichen. Die RaterInnen waren nicht dazu angehalten, separate Punktzahlen für die Dimensionen „Inhalt", „Aufbau" und „Sprache" festzuhalten, sondern lediglich eine finale Punktzahl einzureichen, in der alle drei Dimensionen berücksichtigt waren.

Wie sich an den Werten in Tabelle 2 ablesen lässt, fielen die Interrater-Reliabilitäten sehr gut aus. Entsprechend wurde der Mittelwert aus den drei Urteilen als Maß der Textqualität berechnet.

Projekt / Aufgabe	ICC(C,3)	Mittelwert	Standardabweichung
A / Persuasiver Brief	.94	98.78	14.82
B / Fantasiegeschichte 1	.95	96.98	16.36
B / Fantasiegeschichte 2	.90	98.02	14.52
B / Bastelanleitung 1	.91	99.96	16.11
B / Bastelanleitung 2	.92	100.41	16.69

Tab. 2: Interrater-Reliabilität, Mittelwert und Standardabweichung pro Schreibaufgabe

4 | Diskussion

Dem vorliegenden Artikel lag die Fragestellung zugrunde, welches Textbeurteilungsverfahren sich vor dem Hintergrund des Erkenntnisinteresses, der verfügbaren zeitlichen und finanziellen Ressourcen sowie der forschungsrelevanten Gütekriterien anbietet. Geklärt wurde diese Fragestellung aus Perspektive zweier Dissertationsprojekte, in denen der Erfassung von Textqualität eine zentrale Bedeutung zukommt. Unter Berücksichtigung der drei diskutierten Entscheidungsgrundlagen wurde in den beiden Projekten ein holistisches Benchmarkrating ausgewählt:

Im Hinblick auf das Erkenntnisinteresse steht in beiden Arbeiten die Qualität des Gesamttextes im Zentrum des Interesses. So hat die Textbeurteilung in Projekt A die Funktion einer Fallauswahl, und zwar sollen die schwach schreibenden Jugendlichen identifiziert werden, um in einer weiteren Untersuchung deren Schreibprozess in den Blick zu nehmen. Im Fokus stehen Jugendliche, deren Texte als Ganzes und nicht in Bezug auf spezifische Einzeldimensionen schwach ausfallen. Analog dazu interessiert auch in Projekt B die globale Einschätzung der Texte, um diese mit unterschiedlichen Facetten des schreibbezogenen metakognitiven Wissens in Zusammenhang bringen zu können (vgl. Kapitel 2.1).

Neben dem Erkenntnisinteresse legten auch die zur Verfügung stehenden Ressourcen eine Textbeurteilung in Form eines holistischen Benchmarkratings nahe. Bei beiden Projekten handelt es sich um Qualifikationsarbeiten, in welchen die zeitlichen und finanziellen Ressourcen begrenzt sind. Im Unterschied zu anderen analytischen und holistischen Beurteilungsverfahren wird das holistische Benchmarkrating in der Literatur als vergleichsweise ressourcenschonend beschrieben (vgl. Schoonen 2005, 10; Bouwer/Béguin et al. 2015, 89; vgl. Kapitel 2.2). Analog zu dieser Erfahrung in anderen Arbeiten hat sich auch in den beiden Dissertationsprojekten gezeigt, dass das Verfahren mit vergleichsweise geringem zeitlichem Aufwand angewendet werden kann. So dauerte die Schulung lediglich einen halben Tag, da die Bearbeitung einiger weniger Texte bereits ausreicht, um den Ablauf zu trainieren und ein gemeinsames Verständnis aufzubauen. Des Weiteren wurden die RaterInnen im Verlauf der Textbeurteilung zunehmend schneller und vermochten schließlich innerhalb von fünf bis sieben Minuten ein Globalurteil für einen Text abzugeben. Das Verfahren hat sich folglich nicht nur hinsichtlich der Schulung, sondern auch bezüglich der eigentlichen Durchführung als effizient erwiesen.

Im Zusammenhang mit der Schulung hat sich in den beiden Projekten allerdings eine andere Schwierigkeit herausgestellt: RaterInnen, vor allem Lehrkräfte, sind es in der Regel nicht gewohnt, holistische Beurteilungen vorzunehmen und Benchmarktexte bei ihren Einschätzungen zu berücksichtigen. Vielmehr ist ihnen das Einschätzen einzelner Kriterien vertraut. Hinzu kommt, dass sie häufig ihr Hintergrundwissen heranziehen, was mit unterschiedlichen Erwartungshaltungen an die Textqualität einhergeht. In der oben zitierten Studie B stufte beispielsweise eine Lehrperson beim narrativen Rating Texte mit Gewaltszenen per se schwach ein. Des Weiteren fokussierten einige RaterInnen formale Aspekte stark, unabhängig von der spezifischen Aufgabe und dem darin formulierten Schreibziel. Solche unterschiedlichen Erwartungshaltungen gilt es in den Raterschulungen zu thematisieren, damit diese die Textbeurteilung nicht beeinflussen.

Hinsichtlich des dritten Aspekts, den Gütekriterien, hat der Blick auf die Forschung gezeigt, dass kein Verfahren dem anderen deutlich überlegen ist, sondern mit einer umsichtigen Planung und Durchführung sowohl eine holistische als auch analytische Herangehensweise zufriedenstellende Resultate ermöglicht. Einzig bezüglich der Generalisierbarkeit weist der Stand der Forschung auf einen leichten Vorteil des holistischen Ratings mit Ankertexten hin (vgl. Kapitel 2.3). Die aus den beiden Projekten vorliegenden Daten lassen Aussagen zur Reliabilität, konkret zur Beurteiler-Reliabilität, zu. Die hohen erzielten ICC-Werte für alle eingesetzten Schreibaufgaben manifestieren, dass sich mit dem holistischen Benchmarkrating reliable Textbeurteilungen vornehmen lassen. Ferner lassen sie erkennen, dass das Verfahren über unterschiedliche Anlässe, Textgenres und Altersstufen hinweg robust ist.

Da die Konzipierung einer Skala mit Benchmarktexten mit einem relativ hohen Aufwand verbunden ist, interessiert die Frage, inwiefern sie auf andere Aufgaben und Genres übertragen werden kann. Bouwer/Koster et al. (2016) sind dieser Frage nachgegangen und konnten nachweisen, dass Benchmarkskalen für verschiedene Aufgaben verwendet werden können, solange sie das gleiche Genre betreffen.

Das vorgestellte Verfahren wurde im Forschungskontext konzipiert und hat sich in diesem als geeignet erwiesen. Es stellt sich aber auch die Frage, inwiefern die Benchmarkskalen für den Einsatz in der Schulpraxis geeignet sind bzw. adaptiert werden können. Oberflächlich betrachtet, suggerieren die fünf Benchmarktexte eine Notenskala, wie sie in der Schule eingesetzt wird. Dabei ist jedoch zu beachten, dass der schwächste und stärkste Benchmarktext nicht die Grenzen der Beurteilungsskala darstellen und schlechtere und bessere Texte vorkommen werden. Des Weiteren ist wichtig, dass die Benchmarktexte aus einer für die Altersstufe repräsentativen Stichproben stammen, da die Texte ansonsten für die eigene Klasse zu stark oder zu schwach sein können. Als Instrument zur summativen Beurteilung von Schülertexten ist das Benchmarkrating ohne Anpassungen folglich wenig geeignet. Eine formative Beurteilung lässt sich mit dem vorgestellten Instrument hingegen durchaus vornehmen, wie Sturm/Lindauer et al. (2018) aufzeigen.

Literatur

Böhme, Katrin / Bremerich-Vos, Albert / Robitzsch, Alexander (2009): Aspekte der Kodierung von Schreibaufgaben. In: Granzer, Dietlinde / Köller, Olaf / Bremerich-Vos, Albert / Van den Heuvel-Panhuizen, Marja / Reiss, Kristina / Walther, Gerd (Hg.): Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule. Weinheim: Beltz, S. 290–329.

Bouwer, Renske / Béguin, Anton / Sanders, Ted / Van den Bergh, Huub (2015): Effect of genre on the generalizability of writing scores. In: Language Testing 32/1, S. 83–100.

Bouwer, Renske / Koster, Monika / Van den Bergh, Huub (2016): Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner. Manuscript submitted for publication.

Bundesinstitut bifie (2018): Materialien – Überprüfung der Bildungsstandards. Abrufbar unter: https://www.bifie.at/material/ueberpruefung-der-bildungsstandards/ [7.9.2018].

Canz, Thomas (2015): Validitätsaspekte bei der Messung von Schreibkompetenzen. Dissertation, Humboldt-Universität zu Berlin. Abrufbar unter: https://edoc.hu-berlin.de/dissertationen/canz-thomas-2015-10-19/PDF/canz.pdf [8.9.2018].

Feenstra, Hiske (2014): Assessing writing ability in primary education. On the evaluation of text quality and text complexity. Dissertation, Universiteit Twente. Abrufbar unter: http://doc.utwente.nl/91640/1/thesis_H_Feenstra.pdf [8.9.2018].

Graham, Steve / Harris, Karen / Hebert, Michael (2011): Informing Writing: The Benefits of Formative Assessment. New York: Carnegie Corporation of New York.

Grotjahn, Rüdiger / Kleppin, Karin (2017): Gütekriterien bei der Evaluation von Schreibkompetenzen. In: Akukwe, Bettina / Grotjahn, Rüdiger / Schipolowski, Stefan (Hg.): Schreibkompetenzen in der Fremdsprache: Aufgabengestaltung, kriterienorientierte Bewertung und Feedback. Tübingen: Narr Francke Attempto, S. 41–69.

IQB – Institut zur Qualitätsentwicklung im Bildungswesen (2016): VERA – Ein Überblick. Abrufbar unter: https://www.iqb.hu-berlin.de/vera [7.9.2018].

Konsortium HarmoS Schulsprache (2010): Schulsprache – Wissenschaftlicher Kurzbericht und Kompetenzmodell. Abrufbar unter: http://www.edudoc.ch/static/web/arbeiten/harmos/L1_wissB_25_1_10_d.pdf [8.9.2018].

McGraw, Kenneth O./ Wong, Seok P. (1996): Forming inferences about some intraclass correlation coefficients. In: Psychological Methods 1/1, S. 30–46.

National Assessment Governing Board (2010): Writing Framework for the 2011 National Assessment of educational Progress. Washington, D.C.: National Assessment Governing Board.

Neumann, Astrid (2007): Briefe schreiben in Klasse 9 und 11. Münster: Waxmann.

Neumann, Astrid (2017): Zugänge zur Bestimmung von Textqualität. In: Becker-Mrotzek, Michael / Grabowski, Joachim / Steinhoff, Torsten (Hg.): Forschungshandbuch empirische Schreibdidaktik. Münster: Waxmann, S. 203–219.

Nussbaumer, Markus (1991): Was Texte sind und wie sie sein sollen – Ansätze zu einer sprachwissenschaftlichen Begründung eines Kriterienrasters zur Beurteilung von schriftlichen Schülertexten. Tübingen: Niemeyer.

Rijlaarsdam, Gert / Braaksma, Martine / Couzijn, Michel / Janssen, Tanja / Raedts, Mariet / Van Steendam, Elke / Toorenaar, Anne / Van den Bergh, Huub (2008): Observation of peers in learning to write: Practice and research. In: Journal of Writing Research 1/1, S. 53–83.

Schipolowski, Stefan / Böhme, Katrin (2016): Assessment of writing ability in secondary education: Comparison of analytic and holistic scoring systems for use in large-scale assessments. In: L1 Educational Studies in Language and Literature 16, S. 1–22.

Schneider, Hansjakob / Wiesner, Esther / Lindauer, Thomas / Furger, Julienne (2012): Kinder schreiben auf einer Internetplattform: Resultate aus der Interventionsstudie myMoment2. 0. In: dieS-online 2, S. 1–37.

Schoonen, Rob (2005): Generalizability of writing scores: an application of structural equation modeling. In: Language Testing 22/1, S. 1–30.

Schoonen, Rob (2012): The Validity and Generalizability of Writing Scores: The Effect of Rater, Task and Language. In: Van Steendam, Elke / Tillema, Marion / Rijlaarsdam, Gert / Van den Bergh, Huub (Hg.): Measuring Writing: Recent Insights Into Theory, Methodology and Practices. Leiden: Brill, S. 1–22.

Sturm, Afra / Lindauer, Nadja / Sommer, Tim (2018): Rückmelden – Aufgaben- und lernzielbezogenes Feedback. In: Der Deutschunterricht 70/3, S. 80–91.

Van den Bergh, Huub / De Maeyer, Sven / Van Weijen, Daphne / Tillema, Marion (2012): Generalizability of Text Quality Scores. In: Van Steendam, Elke / Tillema, Marion / Rijlaarsdam, Gert / Van den Bergh, Huub (Hg.): Measuring Writing: Recent Insights Into Theory, Methodology and Practices. Leiden: Brill, S. 23–32.

Weigle, Sara C. (2002): Assessing Writing. Cambridge: Cambridge University Press.

Nadja Lindauer

Zentrum Lesen
Pädagogische Hochschule
der Fachhochschule Nordwestschweiz
nadja.lindauer@fhnw.ch

Tim Sommer

Zentrum Lesen
Pädagogische Hochschule
der Fachhochschule Nordwestschweiz
tim.sommer@fhnw.ch