Zeitschrift: Bulletin suisse de linguistique appliquée / VALS-ASLA

Herausgeber: Vereinigung für Angewandte Linguistik in der Schweiz = Association suisse de linguistique appliquée

Band: - (2016)

Heft: 103: Auf dem Weg zum Text : sprachliches Wissen und Schriftsprachaneignung = Savoir linguistique et acquisition de la littératie = Metalinguistic knowledge and literacy acquisition

Artikel: Ratingverfahren zur Messung von Schreibkompetenz in Schülertexten

Autor: Wilmsmeier, Sabine / Brinkhaus, Moti / Hennecke, Vera

DOI: https://doi.org/10.5169/seals-978658

Ratingverfahren zur Messung von Schreibkompetenz in Schülertexten

Sabine WILMSMEIER
Universität zu Köln, Mercator-Institut für Sprachförderung und Deutsch als Zweitsprache, Triforum, Albertus-Magnus-Platz, D-50923 Köln, Deutschland
sabine.wilmsmeier@mercator.uni-koeln.de

Moti BRINKHAUS
Leibniz Universität Hannover, Institut für Pädagogische Psychologie
Schloßwender Str. 1, D-30159 Hannover, Deutschland
brinkhaus@psychologie.uni-hannover.de

Vera HENNECKE
Leibniz Universität Hannover, Institut für Pädagogische Psychologie
Schloßwender Str. 1, D-30159 Hannover, Deutschland
hennecke@psychologie.uni-hannover.de

This article describes a part of an interdisciplinary research project that examines subcomponents of writing competence and their teachability in an intervention study. The study was conducted in 5th and 9th grades of different school types in Cologne and Hanover, Germany. To find out whether the training of specific subcomponents had an effect on text quality and writing competence in general, the students produced reports and argumentative texts at three different times of measurement. Text quality was then determined by combining two different approaches to text ratings: All texts were rated following a holistic approach and, independently, using an analytical grid. We reflect on the potentials and limitations of our approach to measure text quality and indicate writing competence. The preliminary results emphasize the need for extensive rater trainings and trial runs to ensure the validity and generalizability of the rating grid and thus the conclusions about text quality and writing competence.

Keywords:
assessment of text quality, holistic and analytical rating, rater training

1. Einleitung: Bewertung von Textqualität in der Schule

Im folgendem Beitrag werden zwei unterschiedliche Verfahren zur Bewertung von Schülertexten vorgestellt, die im Rahmen des BMBF-Forschungsprojekts "Teilkomponenten der Schreibkompetenz"[1] entwickelt und eingesetzt wurden, und in ihren jeweiligen Vor- und Nachteilen diskutiert.

Für die Beurteilung von Textqualität und Schreibkompetenz von Schülerinnen und Schülern (SuS) ist in der Schule die Lehrperson zuständig. Bei ihren

---

[1] Das Projekt "Unterrichtliche Förderung von Teilkomponenten der Schreibkompetenz", ein Verbundprojekt der Leibniz Universität Hannover (Prof. Grabowski) und der Universität zu Köln (Prof. Becker-Mrotzek), wird gefördert vom Bundesministerium für Bildung und Forschung (BMBF; Förderkennzeichen: 01GJ1208A und B). Für weitere Informationen s. www.bmbf.schreibkompetenz.com.

Beurteilungen stützen sich die Lehrerinnen und Lehrer in Deutschland auf von der Kultusministerkonferenz (KMK) verabschiedete Bildungsstandards (Klieme et al. 2007), auf die Unterrichtsinhalte sowie auf die Leistung im Klassenvergleich. Qualität und Leistung werden in der Regel in Form einer Schulnote oder einer korrespondierenden Punktezahl bewertet. Die Qualität eines Textes wird im Schulkontext also oft nur anhand einiger weniger inhaltlicher Gesichtspunkte und durch die Erfüllung der Erwartungshaltung der jeweiligen Lehrperson bestimmt. Mithilfe der Textbewertungen aus dem hier herangezogenen Forschungsprojekt soll versucht werden, die Textqualität auf Basis detaillierter inhaltlicher wie sprachlicher Kriterien zu bestimmen und so Aufschluss über die zugrundeliegende Schreibkompetenz der Schreiberinnen und Schreiber zu erhalten.

'Schreiben' wird im Rahmen des Forschungsprojekts im Sinne von Textproduktion verstanden. Ein Text wird primär geschrieben, um zu kommunizieren (Becker-Mrotzek 2014; Becker-Mrotzek & Böttcher 2011). Die vorliegenden Schülertexte bilden so gesehen die schriftlich kommunizierte Antwort auf die vorausgegangene Schreibaufgabe. Die Bearbeitung solcher Schreibaufgaben kann daher als problemlösendes Handeln aufgefasst werden. Nur wenn die SuS die jeweiligen textsortenspezifischen Merkmale kennen und auch selbst in ihren Texten umsetzen können, erfüllen sie die Anforderungen der Lehrperson. Erst dann gilt ihr Text als qualitativ gut, und die SuS können damit gleichzeitig als kompetente Schreibende gelten (Merz-Grötsch 2010). Sie müssen also wissen, was von ihnen erwartet wird, um dieses Wissen auch produktiv für ihren Schreibprozess nutzen zu können.

Zentrale Fragen im Zusammenhang mit Textqualität und Schreibkompetenz gelten der Erforschung von Teilkomponenten der Schreibkompetenz, ihrer Lehr- und Lernbarkeit sowie der Implementierung in der Praxis, um die SuS in ihrem Schreiben zu unterstützen und zu fördern. Im hier berichteten Forschungsprojekt wird Textqualität mithilfe von Ratingverfahren bestimmt und dann als Indikator für die zugrunde liegende Schreibkompetenz der Schreibenden genutzt (Knopp et al. 2014). Ziel dieses Artikels ist es nicht, die Ergebnisse der Interventionsstudie aufzuzeigen. Vielmehr sollen die eingesetzten Verfahren zur Messung von Textqualität vorgestellt und ihre Eignung sowie ihre Grenzen für die Bestimmung von Textqualität im Forschungskontext (und in der Folge ggf. auch im Unterrichtskontext) aufgezeigt werden.

2. Forschungsstand

Es gab in der Vergangenheit einige Studien, die sich mit der Messung von Textqualität mithilfe von Ratingverfahren auseinandergesetzt haben. Ein Großteil dieser Studien geht davon aus, dass nach dem Kompetenz-Performanz-Modell (Weinert 2001) eine gezeigte Leistung Rückschlüsse auf

die zugrunde liegende Kompetenz zulässt. Die vorliegenden Texte gelten also in Bezug auf die Schreibkompetenz der Schreibenden als indikativ.

Mit der Hamburger Aufsatzstudie stand 1987 in Deutschland erstmals die Schreibfähigkeit von SuS der Klasse 11 im Fokus. Ziel war es, anhand von Aufsätzen und Fragebögen Einblick in die Einflussfaktoren von Schreibleistungen und Aufschluss über die Struktur des Aufsatzunterrichts zu erhalten. Jeder Schülertext wurde von zwei Personen (sog. Ratern[2]) bewertet. "Die Überprüfung der Rater ergab, dass objektivere Bewertungen durch Training erzielbar sind" (Hartmann 1989: 97f.). Neuere Veröffentlichungen, die sich mit der Reliabilität der Textbeurteilungen der Hamburger Aufsatzstudie beschäftigen, kommen allerdings zu dem Ergebnis, dass die Übereinstimmungen der Rater nicht als gut oder sehr gut angesehen werden können. Ihre Empfehlung ist es daher, analytische Ratingverfahren mit holistischen zu kombinieren, da man hiermit die Übereinstimmungsrate der Rater steigern könne (Bremerich-Vos & Possmayer 2013).

Ein sehr detailliertes Kriterienraster zur Bewertung von Schülertexten wurde von Nussbaumer & Sieber (1994) mit dem "Zürcher Textanalyseraster" vorgelegt. In späteren Studien, etwa bei Fix & Melenk (2000), Kruse et al. (2012) sowie im vorliegenden Forschungsprojekt, dient es als Basis zur Entwicklung eines eigenen Kriterienrasters zur Textbewertung. So entstand 1998 beispielsweise im Rahmen eines Forschungsprojekts der Pädagogischen Hochschule Ludwigsburg das "Ludwigsburger Aufsatzkorpus". Es setzt sich aus 2 300 Schülertexten der achten Klassenstufe zusammen, die an Haupt- und Realschulen sowie an Gymnasien erhoben wurden. Auch hier wurden die Schülertexte einer Beurteilung der Textqualität unterzogen. Sie wurden jeweils von zwei unabhängigen Lehrern mithilfe eines detaillierten Bewertungsrasters beurteilt, welches auf dem Zürcher Textanalyseraster basiert (Fix & Melenk 2000).

Ähnlich wurde in der Studie "Deutsch-Englisch-Schülerleistungen International" (DESI) im Schuljahr 2003/2004 die Schreibfähigkeit von SuS der Klasse 9 untersucht. Im Fokus der Untersuchung stand die Textsorte Brief (Neumann 2007b). Die Beurteilung der Briefe erfolgte mithilfe von Kodierhandbüchern. Sie enthielten detaillierte Bewertungskriterien, Beispieltexte samt Bewertung (Benchmarktexte), Anweisungen für die Rater und waren für die verschiedenen Messzeitpunkte (MZP) parallel aufgebaut (Neumann 2007a). Bei der Bewertung kamen sowohl dichotome Antwortformate für inhaltliche und formelle Mikromerkmale zum Einsatz als auch Ratingskalen, mit denen globale und sprachlich-textuelle Kategorien bewertet wurden (Neumann 2007a). Des Weiteren weist Neumann (2007a) darauf hin, dass vor allem die Vorbereitung der Auswertungen in Form von

[2] Im Folgenden wird durchgehend lediglich die männliche Form verwendet, wenn sowohl die männliche als auch die weibliche Form gemeint sind.

detaillierten Beschreibungen der zu bewertenden Kategorien und die Raterschulung ausschlaggebend für eine valide Beurteilung der Schülerleistungen seien. Auf Basis der Ergebnisse der Textratings wurden später "Kompetenzstufen für das Schreiben von Briefen" (Neumann 2007b: 74) entwickelt. Allerdings wird die Beurteilung über Kompetenzstufen hinweg kritisch betrachtet. Wie Fix (2007) anmerkt, ist eine eindeutige Zuordnung zu den verschiedenen Kompetenzstufen nur nach einem umfassenden Training möglich. Und selbst dann sei die Zuordnung in Einzelfällen noch problematisch, da nur bestimmte Ausschnitte der Schreibkompetenz erfasst werden, während andere Aspekte außen vor blieben (Fix 2007).

Im Rahmen einer Studie der Deutschen Forschungsgemeinschaft (DFG) zum Thema "Kooperative Schülerrückmeldungen bei der Textüberarbeitung im Deutschunterricht der Grundschule" (KoText) wurden Texte von SuS der Klasse 3 an Grundschulen auf ihre Textqualität hin untersucht. Allerdings liegt hier ein wichtiger Unterschied zu anderen Projekten vor: Von der ermittelten Textqualität sollte nicht auf die zugrunde liegende Schreibkompetenz geschlossen werden, da die Texte lediglich als eine Momentaufnahme der Schreibkompetenz zum jeweiligen Schreibauftrag betrachtet werden können (Kruse et al. 2012). Die Textbewertungen erfolgten durch intensiv geschulte Rater, die zusätzlich ausführliche Manuals mit Definitionen der Antwortformate und Beispieltexte samt Bewertungen erhielten. Kruse et al. (2012) sprechen auch hier wieder die Empfehlung aus, die Schulung gut vorzubereiten, da sich so bereits im Vorfeld die Möglichkeit bietet, Fragen zu den Bewertungskategorien oder ähnlichem auszuräumen und auch die eigentlichen Beurteilungskriterien, wo nötig, anzupassen, um ihre Explizitheit zu verbessern. All diese Erkenntnisse zur Vor- und Aufbereitung der Ratingverfahren (vgl. van Steendam et al. 2012, für eine vertiefte Auseinandersetzung mit Fragen der Messung von Texteigenschaften) flossen in die Konzeption der Ratingverfahren im vorliegenden Projekt mit ein.

3. Studie

Die im Rahmen dieses Artikels vorgestellten Texte und Verfahren zur Messung von Textqualität und damit der zugrunde liegenden Schreibkompetenz entstammen der zweiten Phase eines interdisziplinären Projekts des Bundesministeriums für Bildung und Forschung (BMBF). Zunächst wurden in Phase 1 (2009-2012) Teilfähigkeiten (Adressaten-orientierung, Wortschatz und Herstellung von Kohärenz/Kohäsion) identifiziert, die zu einer allgemeinen Schreibkompetenz beitragen können, und hinsichtlich ihrer Relevanz für verschiedene schulische Texttypen geprüft. In der laufenden Phase 2 mit dem Titel "Unterrichtliche Förderung von Teilkomponenten der Schreibkompetenz" wurden die ermittelten

Teilfähigkeiten in konkrete schreibdidaktische Maßnahmen überführt und ihre Wirksamkeit in einer Interventionsstudie überprüft.

Auf Basis der Ergebnisse der ersten Projektphase wird davon ausgegangen, dass besonders die Teilkomponenten Adressatenorientierung und Kohärenzherstellung textsortenübergreifend für Schreibkompetenz wichtig sind (Knopp et al. 2012, 2013). Ziel der Phase 2 war es, die Wirksamkeit der Schulung dieser Teilkomponenten im Rahmen einer Interventionsstudie zu überprüfen.

3.1 Design der Interventionsstudie

Die Interventionsstudie wurde in den Klasse 5 und 9 an je drei Gesamtschulen und drei Gymnasien in Köln und Hannover durchgeführt. Die Durchführung erfolgte im Kontrollgruppendesign. Pro Schule und Jahrgangsstufe wurden zwei Klassen ausgewählt, von denen nur die eine das Interventionsmaterial erhielt und im Deutschunterricht bearbeitete. Die andere Klasse fungierte als Kontrollgruppe und erhielt während des Interventionszeitraums ausschließlich den regulären Deutschunterricht. Es nahmen also insgesamt zwölf Schulen mit je zwei Klassen an der Erhebung teil. Die Einteilung der Klassen in Interventions- und Kontrollgruppen erfolgte, sofern nicht durch die Auswahl kooperationsbereiter Lehrkräfte bereits vorgegeben, nach dem Zufallsprinzip.

Von jedem Probanden wurden im Vorfeld allgemeine Variablen zu kognitiven und sprachlichen Fähigkeiten wie Arbeitsgedächtnis, Lese- und Schreibflüssigkeit sowie Wortschatz erhoben. Anschließend folgten drei Messzeitpunkte (MZP; als Prätest vor Beginn der Intervention, direkt im Anschluss an die elf-wöchige Intervention und als Follow-Up nach ca. sechs Monaten), in denen die SuS im Klassenverband weitere Aufgaben bearbeiteten. Hierzu gehörten beispielsweise Aufgaben zum Erkennen und Herstellen von Kohärenz mittels Referenzen und Konjunktionen, aber auch Aufgaben zur Perspektivenübernahme (konzeptuell, räumlich-visuell und affektiv-emotional). So kann im Vergleich mit den anfangs vorliegenden Grundvoraussetzungen (gemessen zu MZP 1) ein unmittelbarer (Post-test, MZP 2), aber auch ein möglicher langfristiger Lernzuwachs (Follow-Up, MZP 3) geprüft werden. Zusätzlich zu den bearbeiteten Aufgaben schrieben die SuS zu jedem der drei Messzeitpunkte einen berichtenden und einen argumentativen Text. Die Test- und auch die Schreibaufgaben waren für beide Jahrgangsstufen dieselben. Zu jedem der drei Messzeitpunkte wurden die Schreibaufgaben leicht variiert, jedoch mit dem Ziel, den Schwierigkeitsgrad ähnlich zu halten. Bei den Berichten sollten die Hergänge zweier Unfälle und eines Einbruches geschildert werden, zu denen jeweils Bildimpulse vorlagen. Bei den Argumentationen sollten die Schuldfrage bei einem Unfall, bei einem Missgeschick mit einer Wasserbombe und die Wahl eines von zwei Anschaffungsvorschlägen für die Schule begründet werden. Die gestellten Schreibaufgaben orientieren sich (abgesehen von der Möglichkeit zur direkten

Überprüfung der Leserwirkung) an den von Bachmann & Becker-Mrotzek (2010) vorgeschlagenen Standards für gute Schreibaufgaben. Im Sinne der Objektivität wurden für alle Aufgaben an den drei Messzeitpunkten standardisierte Instruktionen von den Versuchsleitern genutzt (Durchführungsobjektivität).

3.2 Stichprobe und Textkorpus

Insgesamt liegen 589 vollständige Datensätze von SuS der Klasse 5 (305 weibliche und 284 männliche Probanden, davon 47% mit sprachlichem Migrationshintergrund) und 556 Datensätze von SuS der Klasse 9 (291 weibliche und 265 männliche Probanden, davon 41% mit sprachlichem Migrationshintergrund) vor, das sind somit 3 534 Texte aus Klasse 5 und 3 336 Texte aus Klasse 9. Im Sinne der Objektivität (Auswertung & Interpretation) und Validität (vgl. Ingenkamp 1989, zu Reliabilitäts- und Validitätsproblemen bei der Aufsatzbeurteilung) wurden diese handgeschriebenen Texte zunächst transkribiert und orthographisch (aber nicht morpho-syntaktisch) normalisiert, um zu vermeiden, dass Aspekte der Leserlichkeit der Handschrift und der Schriftnorm die funktionale Beurteilung der Texte überdecken.

3.3 Durchführung der angewandten Ratingverfahren

Sowohl holistische als auch analytische Ratingverfahren müssen den Gütekriterien (vgl. z. B. Bortz & Döring 2009) entsprechen, sollen sie mit Erfolg zum Einsatz kommen. Gleiches gilt für die wissenschaftliche Messung von Schulleistungen und Textqualität im Schulbereich (Arnold 2001; Bortz & Döring 2009; Sacher 2014). Während holistische Ratingverfahren über eher allgemeine Aspekte die Textqualität als Ganze (und darüber letztlich auch die Schreibkompetenz) messen sollen, ermöglichen analytische Ratings einen detaillierteren Blick auf einzelne Textelemente und textsortenspezifische Merkmale. Sie messen also in gewisser Weise immer nur eine textsortenspezifische Schreibkompetenz; in diesem Fall für berichtende und argumentative Texte. Unabhängig davon, welches Verfahren man nun anwendet, um Rückschlüsse auf die Textqualität zu ziehen, muss das methodische Vorgehen genau geplant werden. Wie entscheidend bereits die Auswahl und gründliche Vorbereitung der Ratingverfahren ist, wurde bereits in Kapitel 2 angedeutet.

Ebenso weisen auch van den Bergh et al. (2012) darauf hin, dass die Stabilität von Urteilen über Textqualität auch von den angewandten Ratingverfahren abhängig ist. Die Bewertung eines Textes sei dabei nur eine Momentaufnahme der Schreibkompetenz des Schreibenden. Über die Textqualität hinaus könnten keine Rückschlüsse auf die generelle Schreibkompetenz in der untersuchten Textsorte oder textsortenübergreifend gezogen werden (Olinghouse et al. 2012; van den Bergh et al. 2012). Und

dennoch wird genau das in den meisten Studien zu Schülerleistungen getan. Laut van den Bergh et al. (2012) können Ergebnisse aber nur generalisiert werden, wenn mehrere Texte eines Schreibers vorliegen und diese von mehreren Ratern bewertet werden. Zum selben Schluss kommt auch Schoonen (2012) bei der Analyse von englischen Schülertexten mithilfe analytischer Ratings. Er hält fest, dass Verfahren zur Messung von Schreibkompetenz nicht generalisiert werden können, wenn lediglich ein Text pro Proband vorliegt. Da für das vorliegende Projekt aber von jedem Probanden jeweils Texte zu insgesamt sechs verschiedenen Schreibaufgaben vorlagen, gehen wir davon aus, dass die Ergebnisse der Analyse entsprechend zumindest textsortenspezifisch generalisierbar sind und durchaus Rückschlüsse auf die Schreibkompetenz der SuS zulassen.

Um eine möglichst umfassende und aussagekräftige Bewertung der Textqualität zu erhalten, wurden für die Untersuchung der Schülertexte zwei unterschiedliche Verfahren kombiniert – ein holistisches und ein analytisches Ratingverfahren. Dadurch liegen zu jedem Text Mehrfachbeurteilungen vor (Grabowski et al. 2014), was den Empfehlungen und geschilderten Vorgehensweisen vorausgegangener Forschungsprojekte (Neumann 2007a; Kruse et al. 2012; Bremerich-Vos & Possmayer 2013) entspricht und die Objektivität und Reliabilität der Textbewertungen gewährleisten kann bzw. zumindest überprüfbar macht.

3.3.1 Holistisches Rating

Ziel der holistischen Ratings war es, über quantitative Mittel allgemeine Aussagen über die geschriebenen Texte zu erhalten. Zu diesem Zweck wurde ein vergleichsweise einfaches Bewertungsverfahren entwickelt, das sich an allgemeinen Beurteilungsrastern orientiert (z. B. Becker-Mrotzek & Böttcher 2011). Das Bewertungsraster umfasste sechs Fragen nach der *Textqualität* (hoch/ niedrig), der *Erfüllung der Textfunktion* (erfüllt/ nicht erfüllt), dem *Wortschatz* (angemessen/ nicht angemessen), der *Adressatenorientierung* (adressatenorientiert/ nicht adressatenorientiert), der *Kohärenz* (aus sich selbst heraus verständlich/ nicht aus sich selbst heraus verständlich) und dem *inhaltlichen Zusammenhang* (klarer inhaltlicher Zusammenhang/ kein klarer inhaltlicher Zusammenhang). Die Antwortmöglichkeiten waren bewusst dichotom gewählt; in Gegenüberstellung zu detaillierten analytischen Ratings sollte hier ein vergleichsweise einfaches Bewertungssystems zum Einsatz kommen, an dem sich aufwändigere analytische Verfahren in ihrer Effizienz beweisen müssen. Deshalb wurde dieses holistische Rating auch als "naives" Rating durchgeführt, d. h. Studierende des Faches Deutsch (in der Regel Lehramt) beurteilten die Texte ohne spezielle weitere Schulung.

Die Schülertexte wurden in Vorlesungen an Studierende ausgeteilt mit der Bitte, die Texte zu bewerten. Die Studierenden erhielten pro Durchgang jeweils etwa zehn Texte zur Bewertung und hatten dafür mit durchschnittlich

15 Minuten ausreichend Zeit. Den Studierenden wurden die Aufgabenstellungen und Bildimpulse zu den Texten des jeweiligen Messzeitpunktes gezeigt. Anschließend wurde das Bewertungsraster vorgestellt, das unter jedem Text in Form eines Feldes mit sechs Fragen abgebildet war. Die Frage- und Antwortmöglichkeiten waren für beide Textsorten (Berichte und Argumentationen) identisch, um eine gute Vergleichbarkeit der Daten zu gewährleisten. Jeder Text wurde von zwei Ratern bewertet.

Bei den Texten von SuS der Klasse 5 und 9 variierten die Übereinstimmungen der Rater (Inter-Rater-Reliabilität) über die sechs Beurteilungsaspekte und die beiden Textsorten von MZP 1 zwischen ca. knapp 60% und ca. 70%. Interessanterweise erlauben die Texte von MZP 2 und 3 höhere Inter-Rater-Übereinstimmungen, die häufig über 70% liegen, aber in der Regel unter 80% bleiben. Je nach Betrachtungsweise kann man die Höhe dieser Übereinstimmungen als unzureichend ansehen, oder aber angesichts des "naiven", also nicht speziell geschulten Herangehens auch als durchaus noch überzufällig zuverlässig (wenn man an bekannte Befunde zur Aufsatzbeurteilung "echter" Lehrkräfte denkt).

Nachdem sich das jeweils "wahre" dichotome Urteil nicht bestimmen lässt, wurden die Urteile zunächst paarweise pro Urteilsaspekt gemittelt und dann auch über alle sechs Urteilsaspekte aggregiert. Ein derart "robustifizierter" Messwert hat sich im ersten Projektabschnitt als hoch korreliert mit allgemeinen sprachlichen und kognitiven Fähigkeitsvoraussetzungen der SuS erwiesen (Knopp et al. 2012, 2013).

3.3.2 Analytisches Rating

Ein analytisches Rating durchzuführen bedeutet, ein Konstrukt bei der Kodierung sehr fein auszudifferenzieren und in seine Bestandteile zu zerlegen (Behrens & Krelle 2011). Im Rahmen des Projekts geschah dies für die Konstrukte Perspektivenübernahme und Kohärenzherstellung. Perspektivenübernahme bedeutet die Fähigkeit, die wissensbezogene, räumliche und ggf. auch affektiv-emotionale Lage eines Partners zu erkennen und bei der Textproduktion zu berücksichtigen (Schmitt 2011). Unter Kohärenzherstellung verstehen wir die Fähigkeit, Zusammenhänge und Ordnungsstrukturen zu erkennen (im rezeptiven Fall) sowie (bei der Sprach- und Textproduktion) eine Ordnungsstruktur textsortengemäß herzustellen und die zugehörigen Zusammenhänge sprachlich angemessen zu markieren. Für eine Detailanalyse wurden die Aspekte der *Vollständigkeit* der Information (Bericht und Argumentation), der *Lokalisation* der Aktanten (Bericht) bzw. der *sprachlichen Realisierung* von Argumenten (Argumentation) sowie der *Textkohärenz* (Wiederaufnahmen von Beobachter und Aktanten bzw. Argumenten; temporale bzw. logisch-kohäsive Mittel) identifiziert und für die Konstruktion eines Auswertungsschemas mit klaren Merkmalen versehen,

die möglichst alle inhaltlichen Varianten der je drei Zieltexte parallel berücksichtigen (Neumann 2007a). So wird sichergestellt, dass die Bewertungen der Rater einheitlich und objektiv erfolgen. Dadurch war der Aufwand für die analytischen Ratings deutlich höher als für die holistischen Ratings. Es mussten mehrere Testphasen (Pilotierungen) durchlaufen werden, bevor die Test-Gütekriterien (siehe Kapitel 3.3) als erfüllt angesehen werden konnten. Nachdem sich im Prozess der Pilotierung durch Weiterentwicklung der Analyseitems und des Schulungsverfahrens akzeptable Reliabilitäten ergeben haben (im Detail siehe Kapitel 4), wurde die Analyse aller erhobenen Texte dann nur noch von jeweils einem Rater durchgeführt. Die Objektivität der Bewertungen kann entsprechend vorheriger Studien (z. B. Neumann 2007a) zusätzlich durch die Kopplung bzw. Generalaggregation der beiden durchgeführten Ratingverfahren erhöht werden; weiterhin erweist sich auch die bloße Textlänge in der Sekundarstufe I als relevanter Indikator der Textqualität.

Der Ratingprozess für beide Jahrgangsstufen bestand im Kern pro Messzeitpunkt aus einer umfassenden Rater-Schulung, einer nachfolgenden Pilotierung und dem sich anschließenden eigentlichen Rating samt Auswertung. Als Grundlage für die Rating-Items und die Schulung dienten Kriterienraster und Manuals aus der ersten Phase des Projekts, die in Anlehnung an das Zürcher Textanalyseraster von Nussbaumer & Sieber (1994) entwickelt worden waren. Die Items waren inhaltlich auf die Textsorten Bericht und Argumentation und speziell auf die hierzu gestellten Schreibaufgaben zugeschnitten (Knopp et al. 2014; s. oben).

Die Rating-Items gliedern sich jeweils in die oben genannten drei Oberkategorien (*Vollständigkeit, Lokalisation* bzw. *Realisierung* sowie *Textkohärenz*). Die zugehörigen Beurteilungsitems wurden so formuliert, dass sie auf alle drei Aufgabenvarianten der jeweiligen Textsorte anwendbar sind. Die überwiegende Anzahl der Items wurde wiederum dichotom formuliert. Beispielsweise enthielten die Bildstimuli aller drei Schreibaufgaben für Berichte Straßensettings, in denen ein Beobachterstandpunkt definiert war und Aktanten auf der gegenüberliegenden Seite zu benennen und in ihren Handlungen und Bewegungen zu charakterisieren waren. So konnte etwa der Aspekt "Lokalisierung der Aktanten" einheitlich untergliedert werden in die dichotomen Items "Wird der Standpunkt des Berichtenden genannt (Ja/Nein)"; "Wird der Aktant/die Aktanten auf der gegenüberliegenden Straßenseite benannt (Ja/Nein)" und "Erfolgen die Lokalisierungen und Bewegungsbeschreibungen aus der geforderten Beobachtungsperspektive (Ja/Nein)". Analog wurden die Items für die Einschätzung der Textkohärenz und (bei den argumentativen Texten) der sprachlichen Realisierung gebildet. Bei der Kategorie der Vollständigkeit (jeweils untergliedert in die Abfrage der Thematisierung von vier zentralen Handlungselementen) mussten die Items

natürlich spezifisch für das Geschehen in dem jeweiligen Aufgabenstimulus formuliert werden.

Die Rater entscheiden dann jeweils, ob beispielsweise beim Unfallbericht die entscheidende Information, dass das Auto gegen die Straßenlaterne gefahren ist, explizit im Text genannt wird oder nicht. Hinzu kommen zwei Items, bei denen die Rater sprachliche Merkmale auszählen und mittels freier Zifferneingabe eintragen sollten; nämlich die Zahl der verwendeten Lokalisierungen bei den Berichten und die Anzahl der verwendeten temporalen Mittel (im Zusammenhang mit der Kohärenzeinschätzung). Alle Rating-Items blieben nach der zweiten Pilotierung (siehe Kapitel 4.2) im Kern konstant und wurden sowohl für alle drei Messzeitpunkte als auch für beide Jahrgangsstufen angewandt.

In einer mehrstündigen Rater-Schulung wurden der Kontext der Studie, die verschiedenen Aufgabenstellungen sowie die Rating-Items samt Beispielen vorgestellt. Zusätzlich wurden offene Fragen geklärt und einzelne Texte probeweise gemeinsam bewertet. Die Rater erhielten darüber hinaus bereits während der Schulung ein Manual mit den Items, weiterführenden Erklärungen und Beispielen zur Bewertung. Im Anschluss an die Schulung erfolgte dann die erste Pilotierung. Für die Pilotierung wurde eine zufällige Stichprobe von 50 Texten pro Textsorte auf die zehn Rater aufgeteilt, sodass jeder insgesamt jeweils zehn Berichte und zehn Argumentationen zu bewerten hatte.

Für die Durchführung der analytischen Ratings wurde mithilfe der Software Filemaker eine Datenbankstruktur programmiert, in der in einem geteilten Fenster jeweils rechts der zu beurteilende Text und links auf verschiedenen Reiterkarten die einzelnen Beurteilungsitems zu sehen waren; die Ratingeingabe konnte somit durch einfaches Anklicken der Ja/Nein-Felder bzw. Eingabe einer Ziffer erfolgen und über die Rater hinweg bequem zusammengeführt werden. Die Inter-Rater-Reliabilität wurde (siehe oben) in der Pilotierungsphase für die dichotomen Items als prozentuale Übereinstimmung bestimmt; bei den Items mit freier Zifferneingabe wurde die Intra-Klassen-Korrelation (ICC) berechnet. Gemessen an vorliegenden Maßstäben (z. B. Stemler 2004; Graham et al 2012) wurden Übereinstimmungen von mindestens 75% bzw. ICC-Werte ab .7 als ausreichend aufgefasst. In mehreren Schritten wurden diejenigen Items, bei denen diese Werte nicht erreicht werden konnten, in ihrer Formulierung verbessert, im Schulungsmanual besser expliziert oder mit einschlägigeren Beispielen geschult (siehe Kapitel 4.2). Diejenigen Items, bei denen in einer weiteren Pilotierung die kriterialen Werte nicht erreicht werden konnten, wurden für das Hauptrating nicht weiter berücksichtigt.

Insgesamt ergaben sich damit für *Vollständigkeit* sieben Items (Bericht) bzw. fünf Items (Argumentation); für *Lokalisierung* vier Items (Bericht); für *sprachliche Realisierung* sechs Items (Argumentation) und für *Textkohärenz*

fünf Items (Bericht) bzw. vier Items (Argumentation). Nachdem eine hinreichende Reliabilität des letztendlich verwendeten analytischen Ratings sichergestellt war, wurden alle erhobenen Texte nur noch von jeweils einem entsprechend geschulten Rater beurteilt.

In der weiteren – hier nicht mehr berichteten – Analyse wird zunächst anhand der dann erfolgten Ratings geprüft, inwieweit die Items, die zu einer Kategorie gehören, eine Skala bilden; bei hinreichend hoher interner Konsistenz (Cronbach's Alpha) werden sie aggregiert und dann mit den anderen Indikatoren der Textqualität abgeglichen, durch die erhobenen Fähigkeitsmaße der SuS vorhergesagt bzw. im Rahmen des Untersuchungsdesigns (Interventions- vs. Kontrollgruppe; 3 MZP) "signifikanz-kritisch" verglichen.

4. Aufwand und Nutzen der Ratingverfahren

4.1 Holistisches Rating

Die zeitliche Flexibilität der Ratings könnte deutlich verbessert werden (z. B. nur während der Vorlesungszeit), wenn auch die holistischen Ratings onlinebasiert durchgeführt werden würden. Ein weiterer Vorteil dieses Vorgehens bestünde in der verbesserten Transparenz. Der aktuelle Stand der Rückläufe und auch die Auswertung der Ergebnisse könnten ebenfalls zentral und automatisiert erfasst werden. Eine direkte Einbettung solcher Textbewertungen in Lehrveranstaltungen, beispielsweise mithilfe eines Online-Tools, scheint eine gute Möglichkeit zu sein. Für den Fall, dass sich dies nicht realisieren lässt, ist es zumindest sinnvoll, die Auswertung durch den Einsatz einer Scan-Software zu automatisieren. Hierdurch können wichtige Ressourcen wie Arbeitszeit und die Anzahl der Arbeitskräfte effektiver genutzt werden. Dennoch bleibt zu bedenken, dass auch eine Umstellung auf automatisierte Verfahren sicherlich wieder neue Schwierigkeiten mit sich brächte.

Bei dem Layout der Antwortskalen sollte darauf geachtet werden, dass die Antwortmöglichkeiten klar voneinander abgegrenzt sind; vor allem, wenn die Ratings in Papierform durchgeführt werden. Zusätzlich sollte ein Hinweis zum Vorgehen sowohl mündlich erfolgen als auch noch einmal zentral über den Antwortfeldern platziert werden. Ein zusätzlicher, schriftlich platzierter Hinweis könnte helfen, den Folgeaufwand des Aussortierens und der Veranlassung der Neubewertungen der fraglichen Texte zu minimieren.

4.2 Analytische Ratings

Die Vorbereitung und Durchführung der analytischen Ratings war ungleich zeitaufwendiger als bei den holistischen Ratings. Entsprechend den Erkenntnissen aus ähnlichen Forschungsprojekten wurde viel Zeit auf die Erstellung der Items verwendet. Dennoch traten bei der Auswertung der

Ergebnisse für die fünften Klassen noch vermehrt Schwierigkeiten auf, welche die Grenzen dieses Ratingverfahrens für ein Projekt dieser Art aufzeigen und auch die Eignung dieser Verfahren zum Messen von Textqualität und Schreibkompetenz eingrenzen.

Das ursprüngliche Vorhaben, weitgehend textsortenunabhängige Rating-Items zu erstellen, um eine gute Vergleichbarkeit zwischen den beiden Textsorten zu gewährleisten, ließ sich in dieser Form nicht umsetzen. Zwar wurde dieses Vorhaben für Jahrgangsstufe fünf noch durchgeführt und nur der erste Messzeitpunkt geschult und pilotiert. Größtenteils war dies auf die aufgabenspezifischen Unterschiede der Schreibaufgaben zurückzuführen. Daher wurde beschlossen, für die neunten Klassen jeden Messzeitpunkt zu schulen und mit Pilotierungen vorzubereiten. Dieses Vorgehen war zusätzlich wichtig, weil für die Ratings der Jahrgangsstufe neun ein Wechsel innerhalb des Rater-Teams stattfand.[3]

Nach der ersten Pilotierung zeigte sich außerdem, dass einige Items nicht präzise genug formuliert waren. Dies führte zu Verunsicherungen seitens der Rater, wie sie bei einigen Items mit der Bewertung verfahren sollten. Dies spiegelte sich in schlechten Werten der prozentualen Übereinstimmung der beiden Textbewertungen wider (Inter-Rater-Reliabilität). Alle betroffenen Items wurden anhand der Auswertung sowie der in Filemaker mittels Kommentarfunktion erfassten Probleme identifiziert und anschließend überarbeitet. Es handelte sich hauptsächlich um eine sprachliche Präzisierung der Items und die Ergänzung des Manuals um weitere Anschauungsbeispiele samt Wertung. So lag etwa bei Item B.5 der Argumentationstexte ("Werden mindestens zweimal Modalverben verwendet?") die prozentuale Übereinstimmung der beiden Rater nach der ersten Pilotierung lediglich bei 82%. Nach dem Anpassen der Item-Formulierung (Konkretisierung der zu zählenden Modalverben: müssen, wollen, können, sollen, dürfen, mögen/möchten) und einer erneuten Schulung konnte sie jedoch auf 92 % gesteigert werden. Ähnlich wurde mit weiteren Items verfahren, für welche die Rater-Übereinstimmung unter 80% lag oder bei denen Rückfragen zu den Items bestanden. Bei den in Tabelle 1 aufgeführten Items A.2 "Rotes Auto weicht auf die Gegenfahrbahn aus" und A.3 "Rotes Auto stößt mit blauem Auto zusammen" bestand beispielsweise anfangs Unsicherheit darüber, wie explizit die Informationen im Text genannt, beziehungsweise ob auch die Farben der Autos zwingend genannt werden müssen. In der Nachschulung wurde festgelegt, dass etwa der Vorgang des Ausweichens (Item A.2) explizit im Text genannt werden muss. Bezüglich der Autofarbe wurde entschieden,

---

[3] Der Wechsel im Rater-Team war durch den insgesamt langen Durchführungszeitraum der Ratings bedingt. Er sollte von Anfang an mit einkalkuliert und der Aufwand für neue Pilotierungen und Schulungen entsprechend eingeplant werden. Von den zehn ursprünglichen Ratern blieben sieben bis zum Ende der Ratings erhalten, sodass eine vergleichende Analyse der Ergebnisse gut möglich ist (Neumann 2006).

dass sie nicht zwingend genannt werden muss, solange aus der Beschreibung dennoch deutlich hervorgeht, von welchem Auto jeweils die Rede ist. Dies war möglich, da die SuS von einem bestimmten Standpunkt aus über den Unfall berichten sollten. In vielen Texten fanden sich daher Formulierungen wie "...das rechte Auto wollte ausweichen und hat das linke Auto getroffen" oder den Autos wurde bei der ersten Nennung eine Automarke zugewiesen und später wieder aufgegriffen.

Nach einer weiteren ausführlichen Schulung wurde im Rahmen einer zweiten Pilotierung mit einer neuen Stichprobe geprüft, ob die Rater-Übereinstimmung sich verbessert hatte. Da dies der Fall war (vgl. Tabelle 1), konnten anschließend die eigentlichen Ratings durchgeführt werden.

Ein weiterer Störfaktor war das Vorwissen und die damit verbundene Erwartungshaltung der Rater an die verfassten Texte. Jeder Lesende nutzt beim Lesen sein Weltwissen und seine eigene Schreiberfahrung, um den geschriebenen Text zu verstehen und zu bewerten. Da dieser Prozess allerdings immer subjektiv ist, können die Textbewertungen voneinander abweichen. Zwar fließt diese Erwartungshaltung auch bei den naiven Ratings zu einem gewissen Teil mit in die Bewertungen ein. Allerdings gleichen sich unterschiedliche Einschätzungen einzelner globaler Textphänomene durch die Zweifachbewertung eines jeden Textes in der Summe später eher aus. Bei den analytischen Ratings bieten die zahlreichen detaillierten Fragen zum Vorhandensein bestimmter Informationen oder sprachlicher Mittel aber weitaus mehr Möglichkeiten, Texte unterschiedlich zu bewerten. Ist eine Information beispielsweise nur implizit enthalten, kamen unterschiedliche Personen oft auch zu unterschiedlichen Bewertungen. Deshalb wurde versucht, dieser subjektiven Bewertung mit möglichst eindeutigen Rating-Items entgegenzuwirken und die Bewertungen dadurch insgesamt zu objektivieren. So wurde etwa nach der expliziten Nennung bestimmter Informationen gefragt oder die zu zählenden sprachlichen Mittel wurden exemplarisch aufgelistet. Dies war zusätzlich wichtig, da besonders die Verwendung der in der Intervention geschulten sprachlichen Mittel (bspw. Modalverben und temporale Mittel) interessant war.

Item Nr.	Item – MZP 1 (Unfallbericht)	Klasse 5 1. Pilotierung	Klasse 5 2. Pilotierung	Klasse 9 Pilotierung
A.2	Rotes Auto weicht auf die Gegenfahrbahn aus.	70%	78%	84%
A.3	Rotes Auto stößt mit blauem Auto zusammen.	72%	88%	96%
C.4	Wie viele temporale Mittel werden verwendet?	.59	.73	.80

Tabelle 1: Ausgewählte Reliabilitäten bei der Pilotierungen der analytischen Ratings (Bericht); angegeben sind prozentuale Übereinstimmungen (A.2 und A.3) bzw. Intra-Klassen-Koeffizienten (C.4).

Tabelle 1 verdeutlicht, dass für MZP 1 sowohl die prozentuale Übereinstimmung (Item Nr. A.2 und A.3) als auch die Intra-Klassen-Korrelation (Item Nr. C.4) der Rater für Klasse 9 von Anfang an deutlich höher liegen als für die Texte aus Klasse 5. Somit waren keine Nachschulungen oder weitere Pilotierungen nötig. Dies deutet darauf hin, dass die Maßnahmen (Schulungen zu allen drei Schreibaufgaben, präzisierte Items, ausführlichere Manuals) Wirkung zeigen. Und auch die Tatsache, dass sieben der ursprünglichen zehn Rater bis zum Ende beibehalten werden konnten, spielt eine große Rolle. Zwar konnten durch das Besprechen weiterer Beispiele und den Austausch während der Schulungen zu den drei Messzeitpunkten letzte Zweifelsfälle ausgiebig diskutiert werden. Dennoch waren die Rater bereits merklich sicherer in ihrem Vorgehen, da sie mit dem Vorgehen und den Items schon Erfahrung hatten. Die Möglichkeit, Fragen an die gemeinsame Mailingliste zu stellen, wurde für die Texte aus Klasse 9 stärker genutzt. Alle Teilnehmerinnen und Teilnehmer profitieren von der Diskussion. Dieses Vorgehen ist u. E. empfehlenswert und trägt dazu bei, Unsicherheiten seitens des Rater-Teams abzubauen und damit die Rater-Übereinstimmung zu verbessern. Es bietet auch die Gelegenheit, die eigenen Bewertungen immer wieder mit den Vorgaben abzugleichen. Die Rater laufen so nicht Gefahr, in eine Art Routine zu verfallen, sondern hinterfragen ihre Entscheidungen fortwährend kritisch.

5. Fazit

Verfahren zur Messung von Textqualität sind unterschiedlich aufwändig im empirischen Einsatz und unterschiedlich zuverlässig. (Über ihre Validität kann man in Ermangelung geeigneter Kriterien oft keine Aussage machen.) Sowohl die Handhabung als auch die Messgüte holistischer wie analytischer Ratings kann man aber beeinflussen und verbessern. Nach den berichteten

Erfahrungen im Rahmen einer umfangreichen Interventionsstudie halten wir vor allem die nachfolgenden Aspekte für beachtenswert.

Beim Einsatz holistischer Ratings empfiehlt es sich, vorab abzuwägen, sie onlinebasiert durchzuführen oder zumindest den Auswertungsprozess zu automatisieren. Dies erfordert zwar eine detaillierte Planung in Bezug auf die technischen Voraussetzungen, die praktische Umsetzung und auch die anschließende Auswertung. Dafür profitiert man allerdings von einem logistisch deutlich vereinfachten Prozess. Nach dem bisherigen Stand der Auswertungen und auch aufgrund der Erfahrungen aus der ersten Projektphase gehen wir davon aus, dass die holistischen Ratings sich gut eigenen, um Aussagen über die Textqualität allgemein und damit auch die Schreibkompetenz der SuS zu treffen. Sie sind robust, und ihre Erhebung ist zwar vergleichsweise "quick", aber von der Indikationskraft her nicht "dirty".

Bei der Durchführung der analytischen Ratings hat sich gezeigt, dass das Rater-Training und die sorgfältige Formulierung der Rating-Items sowie ihre Pilotierungen für die spätere Auswertung und Reliabilität der Ergebnisse von Bedeutung sind. Durch die schrittweise Verbesserung der Item-Formulierung (Explizitheit), die Ergänzung der Manuals um weitere Beispiele und Schulungen zu allen drei Messzeitpunkten konnte die Übereinstimmungen der Rater-Urteile deutlich verbessert werden (siehe Kapitel 4.2). Trotz des deutlich höheren Arbeits- und Zeitaufwandes bieten die analytischen Ratings die Möglichkeit, das Konstrukt Textqualität zu zerlegen und einzelne Aspekte separat sowie in Verbindung mit den Ergebnissen der naiven Ratings zu untersuchen.

Die beiden Ratingverfahren sind jedes für sich geeignet, um bestimmte Aspekte von Textqualität zu untersuchen. Die holistischen Ratings beleuchten den Gesamteindruck, den der Text beim Lesenden hinterlässt, indem Kategorien wie Funktionalität und Angemessenheit des Wortschatzes mit berücksichtigt werden. Die analytischen Ratings ermöglichen hingegen einen detaillierten Blick auf einzelne Aspekte der Schülertexte, sind allerdings auch mit einem deutlich höheren Aufwand verbunden. Je nach Fragestellung scheinen beide Verfahren geeignet, um Textqualität zu messen. Allerdings muss der jeweilige Aufwand der Methode mit dem Nutzen in Relation gesetzt werden.

LITERATUR

Arnold, K.-H. (2001). Qualitätskriterien für die standardisierte Messung von Schülerleistung. In F. Weinert (Hg.), *Leistungsmessungen in Schulen*, 2. Auflage (S. 117-130). Weinheim/ Basel: Beltz.

Bachmann, T. & Becker-Mrotzek, M. (2010). Schreibaufgaben situieren und profilieren. In T. Pohl & T. Steinhoff (Hgg.), *Textformen als Lernformen. Kölner Beiträge zur Schreibforschung* (S. 191-210). Duisburg: Gilles & Francke.

Becker-Mrotzek, M. (2014). Schreibkompetenz. In J. Grabowski (Hg.), *Sinn und Unsinn von Kompetenzen: Fähigkeitskonzepte im Bereich von Sprache, Medien und Kultur* (S. 51-71). Opladen: Budrich.

Becker-Mrotzek, M. & Böttcher, I. (2011). *Schreibkompetenz entwickeln und beurteilen. Praxishandbuch für die Sekundarstufe I und II,* 3. Auflage. Berlin: Cornelsen.

Behrens, U. & Krelle, M. (2011). Schülertexte beurteilen im Licht von Bildungsstandards, Kompetenzrastern und Unterrichtsalltag. *Bulletin suisse de linguistique appliquée, 94*, 167-183.

Bergh, H. van den, De Mayer, S., van Weijen, D. & Tillema, M. (2012). Generalizability of Text Quality Scores. In E. van Steendam, M. Tillema, G. Rijmaarsam & H. van den Bergh (Hgg.*), Measuring writing: Recent insights into theory, methodology and practices* (S. 23-32). Leiden: Brill.

Bremerich-Vos, A. & Possmayer, M. (2013). Zur Überprüfung eines textsortenübergreifenden Modells der Entwicklung von Schreibkompetenz in der Grundschule. In A. Redder & S. Weinert (Hgg.), *Sprachförderung und Sprachdiagnostik – interdisziplinäre Perspektiven* (S. 277-295). Münster: Waxmann.

Bortz, J. & Döring, N. (2009). *Forschungsmethoden und Evaluation*. 4. überarbeitete Auflage. Heidelberg: Springer.

Fix, M. & Melenk, H. (2000). *Schreiben zu Texten - Schreiben zu Bildimpulsen. Das Ludwigsburger Aufsatzkorpus*. Baltmannsweiler: Schneider.

Fix, M. (2007). Zur Problematik von Kompetenzstufen für Schülertexte. In H. Willenberg (Hg.), *Kompetenzhandbuch für den Deutschunterricht. Auf der empirischen Basis des DESI-Projekts* (S. 84-95). Baltmannsweiler: Schneider.

Grabowski, J., Becker-Mrotzek, M., Knopp, M., Jost, J. & Weinzierl, C. (2014). Comparing and combining different approaches to the assessment of text quality. In D. Knorr, C. Heine & J. Engberg (Hgg.), *Methods in writing process research* (S. 147-165). Frankfurt am Main: Lang.

Graham, M., Milanowski, A. & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Center for Educator Compensation Reform: Westat.

Hartmann, W. (1989). Die Hamburger Aufsatzstudie. *Der Deutschunterricht, 41*, 3, 92-98.

Ingenkamp, K. (1989). *Diagnostik in der Schule*. Weinheim/ Basel: Beltz

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.E. & Vollmer, H.J. (2007). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Bonn/ Berlin: BMBF (Bildungsforschung, Band 1).

Knopp, M., Jost, J., Nachtwei, N., Becker-Mrotzek, M. & Grabowski, J. (2012). Teilkomponenten von Schreibkompetenz untersuchen: Bericht aus einem interdisziplinären empirischen Projekt. In H. Bayrhuber et al. (Hgg.), *Formate Fachdidaktischer Forschung: Empirische Projekte – historische Analysen – theoretische Grundlegungen* (S. 47-66). Münster: Waxmann.

Knopp, M., Becker-Mrotzek, M. & Grabowski, J. (2013). Diagnose und Förderung von Teilkomponenten der Schreibkompetenz. In A. Redder & S. Weinert (Hgg.), *Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven* (S. 296-315). Münster: Waxmann.

Knopp, M., Jost, J., Linnemann, M. & Becker-Mrotzek, M. (2014). Textprozeduren als Indikatoren von Schreibkompetenz – ein empirischer Zugriff. In T. Bachmann & H. Feilke (Hgg.), *Werkzeuge des Schreibens – Beiträge zu einer Didaktik der Textprozeduren* (S. 111-128). Stuttgart: Fillibach bei Klett.

Kruse, N., Reichardt, A., Herrmann, M., Heinzel, F. & Lipowsky, F. (2012). Zur Qualität von Kindertexten. Entwicklung eines Bewertungsinstruments in der Grundschule. *Didaktik Deutsch, 32*, 87-110.

Merz-Grötsch, J. (2010). *Texte Schreiben lernen. Grundlagen, Methoden, Unterrichtsvorschläge*. 1. Auflage. Seelze: Kallmeyer.

Neumann, A. (2006). Stabilität von Raterurteilen über die Zeit, Anpassung an vorhandene Schülerleistungen: Auswertung zweier Replikationsstudien zu den Urteilen in "DESI-Textproduktion". *Empirische Pädagogik*, *20, 3*, 286-296.

Neumann, A. (2007a). *Briefe schreiben in Klasse 9 und 11. Beurteilungskriterien, Messungen, Textstrukturen und Schülerleistungen*. Münster: Waxmann (Empirische Erziehungswissenschaft, 4).

Neumann, A. (2007b). Schreiben. Ausgangspunkt für eine kriteriengeleitete Ausbildung in der Schule. In H. Willenberg (Hg.), *Kompetenzhandbuch für den Deutschunterricht. Auf der empirischen Basis des DESI-Projekts* (S. 74-83). Baltmannsweiler: Schneider.

Nussbaumer, M. & Sieber, P. (1994). Texte analysieren mit dem Zürcher Textanalyseraster. In P. Sieber (Hg.), *Sprachfähigkeiten – Besser als ihr Ruf und nötiger denn je! Ergebnisse und Folgerungen aus einem* Forschungsprojekt (S. 141-186). Aarau: Sauerländer (Reihe Sprachlandschaft, Bd. 12).

Olinghouse, N. G., Santangelo, T. & Wilson, J. (2012). Examining the validity of single-occasion, single-genre, holistically scored writing assessments. In E. van Steendam, M. Tillema, G. Rijmaarsam & H. van den Bergh (Hgg.), *Measuring writing: recent insights into theory, methodology and practices* (S. 55-82). Leiden: Brill.

Sacher, W. (2014). *Leistungen entwickeln, überprüfen und beurteilen. Bewährte und neue Wege für die Primar- und Sekundarstufe* (5. überarbeitete und erweiterte Auflage). Bad Heilbrunn: Klinkhardt.

Schmitt, M. (2011). *Perspektivisches Denken als Voraussetzung für adressatenorientiertes Schreiben*. Dissertation, Pädagogische Hochschule Heidelberg. Online http://opus.ph-heidelberg.de /frontdoor/index/index/docId/35 (zuletzt abgerufen am 20.04.2016).

Schoonen, R. (2012). The validity and generalizability of writing scores: the effect of rater, task and language. In E. van Steendam, M. Tillema, G. Rijaarsdam & H. van den Bergh (Hgg.), *Measuring Writing: recent insights into theory, methodology and practices* (S. 1-22). Leiden: Brill.

Steendam, E. van, Tillema, M., Rijlaarsdam, G. & van den Bergh, H. (2012). *Measuring Writing: recent insights into theory, methodology and practices*. Leiden: Brill.

Stemler, S. E. (2004). *A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation*, 9, 4.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Hgg.), *Defining and selecting key competencies* (S. 45-65). Göttingen: Hogrefe.