Universität Duisburg-Essen
Fakultät für Geisteswissenschaften
Institut für Germanistik
Modul: Schule und Unterricht forschend verstehen
Seminar: Fachdidaktisches Begleitseminar
Dozentin: Dr. Ulrike Behrens
Wintersemester 2019/2020

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

Einfluss holistischer und analytischer Beurteilungsverfahren auf die Leistungsbewertung im Fach Deutsch

Eine empirische Untersuchung

Sara Cremer
Abgabedatum: 11. März 2020

Inhaltsverzeichnis

1 Einleitung

Eine gerechte Benotung von Aufsätzen stellt vor allem für Lehrkräfte mit geringer Erfahrung im Fach Deutsch immer wieder eine Herausforderung dar und führt sowohl auf Seiten der Lehrenden als auch auf Seiten der Schüler*innen schnell zu Frustration. Ein Grund dafür ist, dass die Beurteilungen vor allem bei mittelmäßigen Leistungen nicht selten divergieren und die Noten aufgrund dessen unter Umständen willkürlich erscheinen (vgl. Birkel & Birkel, 2002).

Um eine möglichst faire und transparente Aufsatzbenotung zu ermöglichen, werden in schulischen Kontexten häufig Kriterienkataloge eingesetzt. Diese sind den analytischen Beurteilungsverfahren zuzuordnen, bei denen Einzelmerkmale eines Textes beurteilt und anschließend zu einer Gesamtnote zusammengefasst werden. Davon zu unterscheiden sind holistische Verfahren, die dadurch gekennzeichnet sind, dass das gesamte Textprodukt als ein Ganzes betrachtet und mit einer einzigen Note beurteilt wird. Auch Beurteilungsverfahren dieser Art sind in der Schule verbreitet, denn sie sind zeitsparend und für erfahrene Lehrkräfte einfach zu handhaben.

Da beide Arten der Leistungsbeurteilung in der schulischen Praxis Anwendung finden, stellt sich in diesem Kontext offensichtlich die Frage, wie stark sich holistische und analytische Beurteilungsverfahren im Hinblick auf die vergebenen Noten unterscheiden. Diese Fragestellung zu untersuchen, ist das Ziel dieses Forschungsprojekts. Dazu wurden die Schreibprodukte von Schülerinnen und Schülern der 5. Jahrgangsstufe sowohl mit holistischen als auch mit analytischen Verfahren beurteilt und die Noten anschließend im Hinblick auf ihre Verteilung und Streuung verglichen.

In Kapitel 2 werden zunächst die Gründe für die Komplexität der Aufsatzbeurteilung sowie eine Auswahl von Verfahren der Leistungsbeurteilung beschrieben. Anschließend werden in Kapitel 3 die Stichprobe, das Design sowie die Methoden der Auswertung beschrieben. Die Ergebnisse des Forschungsprojekts werden in Kapitel 4 zunächst vorgestellt und daraufhin in Kapitel 5 diskutiert und auf die formulierten Hypothesen bezogen. In Kapitel 6 sollen die Ergebnisse des Projekts abschließend zusammengefasst und weitere interessante Forschungsfragen aufgezeigt werden.

2 Theoretischer Hintergrund

2.1 Komplexität der Aufsatzbeurteilung

Die Leistungsbeurteilung ist eine der wesentlichen Tätigkeiten einer Lehrkraft in schulischen Kontexten und stellt zugleich ein höchst komplexes und nicht zu unterschätzendes Unterfangen dar. Als besonders problematisch wird in diesem Zusammenhang die Aufsatzbeurteilung wahrgenommen. Einerseits äußern Schülerinnen und Schüler Unzufriedenheit über eine intransparente und willkürliche Notenvergabe, andererseits beklagen Lehrkräfte die Schwierigkeit der Aufgabe und den großen Arbeitsaufwand (vgl. Grausam, 2018, S. 175).

Ein Grund für die hohe Komplexität der Aufsatzbeurteilung liegt in der Tatsache, dass Textqualität kein „dem Text innewohnendes Merkmal ist“ (Becker-Mrotzek, 2008, S. 175), sondern zahlreiche sprachliche und nicht-sprachliche Kriterien zur Qualität eines Textes beitragen. Daher kann Textqualität, welche häufig als Indikator für Schreibkompetenz herangezogen wird, nicht umstandslos im sozialwissenschaftlichen Sinne gemessen werden, was dazu führt, dass es sich bei den vergebenen Noten bestenfalls um Schätzungen handeln kann (vgl. Becker-Mrotzek & Böttcher, 2015, S. 122).

Eine Replikationsstudie von Birkel zeigt in diesem Zusammenhang, dass die Benotung ein und desselben Textprodukts durch unterschiedliche Lehrkräfte immer noch beinahe die gesamte Notenskala umfasst. Die Streuung ist vor allem bei mittelmäßigen Leistungen groß, sodass Beurteilungen lediglich genutzt werden können, um eine Reihenfolge innerhalb der Klasse zu erzeugen (vgl. Birkel & Birkel, 2002, S. 222f.). Ein Grund für die Schwierigkeit der Beurteilung von mittelmäßigen Textprodukten ist, dass sie sowohl positive als auch negative Aspekte aufweisen. Lehrkräfte neigen in diesem Fall dazu Kriterien heranzuziehen, die leicht zu bewerten sind, aber nicht solche, die Einfluss auf die Textqualität haben (vgl. Becker-Mrotzek, 2008, S. 175).

Diese Unsicherheit bei der Beurteilung von Textprodukten der Lehrerinnen und Lehrer ist verständlich, wenn man beachtet, dass die Fähigkeit zum Beurteilen in den Ausbildungsphasen nur unzureichend thematisiert wird und Lehrkräfte ihre Beurteilungsverfahren daher funktional während des Berufslebens selbst entwickeln. Aus diesem

Grund lässt sich eine Tendenz zur Mitte erkennen, welche nicht nur die Unsicherheit der Lehrkräfte widerspiegelt, sondern auch zeigt, dass sie um die Subjektivität ihrer Beurteilungen wissen und daher extreme Urteile vermeiden (vgl. Grausam, 2018, S. 176).

Um eine faire Leistungsbeurteilung in schulischen Kontexten zu ermöglichen, ist es unerlässlich, dass die gewählten Beurteilungsverfahren die Gütekriterien *Validität, Reliabilität* und *Objektivität* in einem möglichst hohen Maß erfüllen (vgl. Baurmann, 2006, S. 126). Die Validität einer Beurteilung zeigt sich darin, dass mit der gewählten Schreibaufgabe nur die zu beurteilende Leistung gemessen wird und nicht relevante Aspekte dementsprechend vernachlässigt werden. In unterschiedlichen Studien konnte allerdings nachgewiesen werden, dass sowohl Textlänge als auch Handschrift einen signifikanten Einfluss auf die vergebene Note haben und die Validität aufgrund dessen nur eingeschränkt gegeben sein kann. Ähnliche Ergebnisse lassen sich hinsichtlich der Reliabilität finden, da sowohl Schwankungen bei der mehrfachen Beurteilung durch einen Rater als auch Unterschiede in der Beurteilung zwischen verschiedenen Ratern festgestellt werden konnten. Darüber hinaus zeigen Rater mit längerer Erfahrung eine höhere Reliabilität gegenüber Ratern mit geringer Routine. Besonders fehleranfällig ist jedoch die Objektivität, da die Beurteilung von Textprodukten von Schüler*innen nicht von persönlichen Einflüssen loszulösen ist (vgl. Grausam, 2018, S. 180f.).

Aus diesem Grund stellt sich die Frage, ob die Beurteilung mithilfe von Noten im Rahmen der Aufsatzbewertung überhaupt sinnvoll ist. Allerdings ist die Vergabe von Ziffernoten in schulischen Kontexten zurzeit üblich und auf absehbare Zeit nicht überwindbar, sodass auch die Schreibdidaktik nicht auf eine Beurteilung von Textprodukten mithilfe von Noten verzichten kann (vgl. Becker-Mrotzek & Böttcher, 2015, S. 123).

2.2 Verfahren der Leistungsbeurteilung in der schulischen Praxis

Um trotz der dargestellten Schwierigkeiten eine möglichst faire Leistungsbeurteilung von Textprodukten in der Schule zu ermöglichen, haben sich eine Reihe von Verfahren etabliert, die sich in ihrem Aufwand und unter Umständen auch in ihrer Zielsetzung unterscheiden. Baurmann nennt in diesem Kontext die Mehrfachbewertung nach globalem Ersteindruck, den Vergleich anhand einer Textkollektion und den Einsatz von Kriterienkatalogen (vgl. Baurmann, 2006, S. 127ff.).

Diese verschiedenen Verfahren der Leistungsbeurteilung lassen sich ganz grundsätzlich mithilfe der Kategorien *holistisch* und *analytisch* unterscheiden. Bei einer analytischen Beurteilung wird eine relativ große Anzahl an Variablen verwendet, um einzelne Aspekte eines Textprodukts beurteilen zu können. Verfahren dieser Art versprechen eine objektivere und deswegen zuverlässigere Beurteilung, da die Rater kriteriengeleitet vorgehen müssen. Darüber hinaus können die Kompetenzen der Schüler*innen hinsichtlich bestimmter Aspekte der Schreibkompetenz besser beschrieben und für die zukünftige Förderung genutzt werden. Aus diesem Grund eignet sich eine analytische Beurteilung vor allem zu Diagnosezwecken, um Stärken und womögliche Schwächen der Lernenden differenzierter betrachten und untersuchen zu können. Allerdings sind diese Verfahren relativ zeitaufwendig, da das vorliegende Textprodukt immer wieder im Hinblick auf unterschiedliche Kriterien untersucht und beurteilt werden muss (vgl. Schipolowski & Böhme, 2016, S. 2). Des Weiteren muss die Komplexität betont werden, die es bedeutet, Kriterien für die Beurteilung eines Textprodukts aufzustellen (vgl. Bacha, 2001, S. 372).

Ein analytisches Beurteilungsverfahren, welches im Schulalltag weit verbreitet ist und darüber hinaus als äußerst verlässlich gilt, ist der Einsatz von Kriterienkatalogen. Durch die Verwendung kann der Wunsch der Schülerinnen und Schüler nach Transparenz nachgekommen werden und die Aspekte, nach denen das Textprodukt benotet wurde, werden explizit konkretisiert. Diese Tatsache ist nicht allein für die Produktion und die Beurteilung eines Textes sondern vor allem für die Überarbeitung von großer Relevanz. Nur wenn die Lernenden wissen, wie das fertige Textprodukt aussehen soll, sind sie

in der Lage, einen Text dementsprechend zu produzieren (vgl. Becker-Mrotzek & Böttcher, 2015, S. 123).

Das im didaktischen Forschungskontext entstandene und in der Forschung bewährte Zürcher Textanalyseraster (vgl. Nussbaumer & Sieber, 1994, S. 153ff.) bietet eine umfangreiche Grundlage für die Entwicklung von Kriterienkatalogen, muss aber aufgrund seiner Komplexität an die schulische Praxis angepasst werden (vgl. Baurmann, 2006, S. 132). In Anlehnung daran haben Becker-Mrotzek und Böttcher einen Basiskatalog entwickelt, der zwölf Kriterien auf die fünf Basiskategorien *Sprachrichtigkeit, Sprachangemessenheit, Inhalt, Aufbau* und *Prozess* aufteilt. Diese Kriterien können theoretisch für alle Textsorten verwendet werden, wobei sie natürlich an die jeweilige Textsorte sowie die gestellte Schreibaufgabe angepasst werden müssen (vgl. Becker-Mrotzek & Böttcher, 2015, S. 129ff.).

Je weniger Variablen zur Beurteilung verwendet werden, desto ähnlicher ist ein analytisches Beurteilungsverfahren einer holistischen Beurteilung. Dabei wird das Textprodukt als Ganzes betrachtet und die abschließende Bewertung in einem einzigen Urteil zusammengefasst. Dieses Vorgehen hat zur Folge, dass ein Rater gezwungen ist, sprachliche, strukturelle und inhaltliche Eindrücke in einem einzigen Urteil zusammenzufassen. Der bedeutende Vorteil an dieser Art der Beurteilung ist die Zeiteffizienz, da das Textprodukt als Ganzes oder anhand weniger Kriterien benotet wird. Allerdings ist die Beurteilung eher undifferenziert und aufgrund dessen kaum hilfreich für eine Diagnose der Schreibkompetenz der Lernenden (vgl. Schipolowski & Böhme, 2016, S. 3).

Wird in der schulischen Praxis auf holistische Verfahren zurückgegriffen, so geschieht dies meist, indem eine Lehrkraft eine Note vergibt, ohne auf ausformulierte Kriterien zurückzugreifen. Will man die diesem Verfahren zugrundeliegende Subjektivität relativieren, eignet sich die Mehrfachbeurteilung nach globalem Ersteindruck. Dabei wird ein Aufsatz von drei bis vier Ratern unabhängig voneinander beurteilt und die entstehenden Noten anschließend gemittelt, um eine durchschnittliche Note für jedes Textprodukt zu erhalten. Allerdings ist die Erfahrung einer Lehrkraft ein in diesem Kontext nicht zu unterschätzendes Kriterium, da eine gewisse Erfahrung mit Referenzkorpora in diesem Zusammenhang unumgänglich ist, um die Produkte von Schüler*innen sinnvoll

einschätzen zu können (vgl. BAURMANN, 2006, S. 127ff.).
In Studien konnte nachgewiesen werden, dass sowohl die Verwendung von Kriterienkatalogen als auch die Mehrfachbewertung nach globalem Ersteindruck das Kriterium der Reliabilität hinreichend erfüllen. Hinsichtlich der Objektivität führte die Mehrfachbewertung nach globalen Ersteindruck sogar zu besseren Werten als die Verwendung von Kriterienkatalogen. Aus schulpraktischer Sicht können aber natürlich nur gelegentlich Teams für eine globale Mehrfachbewertung gebildet werden, da es sich um ein ressourcen- und zeitaufwendiges Verfahren handelt (vgl. DEHN & BAURMANN, 2004, S. 9f.).
Es lässt sich festhalten, dass beide Formen der Leistungsbewertung ihre Daseinsberechtigung haben und ihr Einsatz in unterschiedlichen Kontexten hilfreich und sinnvoll sein kann. Allerdings sollten beide Verfahren zuverlässige Indikatoren von Schreibkompetenz sein und zu vergleichbaren Beurteilungen führen, um eine faire Benotung in schulischen Kontexten zu ermöglichen. Es stellt sich infolgedessen die Frage, ob holistische und analytische Verfahren der Leistungsbeurteilung zu vergleichbaren Ergebnissen führen. Falls sich Differenzen nachweisen lassen, ergibt sich darüber hinaus selbstverständlich die Frage, in welchen Teilbereichen die Unterschiede besonders groß sind. Als Teilbereiche werden in diesem Kontext *Sprache, Inhalt* und *Aufbau* verstanden.
Aus diesen Überlegungen können folgende Forschungshypothesen abgeleitet werden, die im Verlauf der Forschungsprojekts untersucht werden sollen:

- Die durch analytische und holistische Verfahren vergebenen Noten unterscheiden sich maßgeblich.
- Vor allem im Mittelfeld differieren die Beurteilungen stark und fokussieren sich auf einzelne Teilbereiche.
- Im Rahmen der holistischen Beurteilung hat der Teilbereich *Sprache* einen besonders starken Einfluss auf die Gesamtnote.

3 Methode

3.1 Stichprobe

Um die Fragestellung zu untersuchen, wurden die Textprodukte von Schülerinnen und Schüler einer 5. Klasse einerseits mit holistischen Verfahren und andererseits mit analytischen Verfahren der Leistungsbeurteilung bewertet.
Bei der Stichprobe handelt es sich um 23 Schüler*innen der 5. Klasse eines Gymnasiums, wobei sich die Klasse aus 14 Mädchen und neun Jungen zusammensetzt. Die Klasse hat drei Stunden Deutschunterricht in der Woche und zusätzlich steht eine Stunde Freiarbeit für das Fach Deutsch zur Verfügung.

3.2 Design

Die Schülerinnen und Schüler erhielten im Anschluss an die Unterrichtsreihe zum Thema *Bildergeschichten* eine Schreibaufgabe, die eine vergleichbare Aufgabenstellung wie die Klassenarbeit enthielt (siehe Anhang). Während der Unterrichtsreihe wurde gemeinsam mit den Schüler*innen ein Kriterienkatalog für die Beurteilung von Bildergeschichten entwickelt. Dieser enthält größtenteils Kriterien, welche die Lehrkraft bereits am Anfang der Unterrichtsreihe mit den Lernenden herausgearbeitet hat.
Darüber hinaus ist der Kriterienkatalog an den von Becker-Mrotzek und Böttcher entwickelten Basiskatalog angelehnt. Bei der Entwicklung wurde sich vorrangig an dem Katalog orientiert, den Becker-Mrotzek und Böttcher für das Schreiben von komplexen Geschichten in der Grundschule und der 5. Klasse vorschlagen, da dieser den von der Klasse entwickelten Kriterien am meisten entsprach (vgl. Becker-Mrotzek & Böttcher, 2009, S. 64). Der vollständige Kriterienkatalog enthält 13 Kriterien und umfasst die Kategorien *Sprachrichtigkeit, Sprachangemessenheit, Inhalt, Aufbau* und *Prozess* (siehe Anhang). Die Schreibaufgabe wurde in einem für die Lernenden üblichen Klassenraumsetting durchgeführt. Den Schülerinnen und Schülern stand eine Zeitstunde für die Bearbeitung der Schreibaufgabe zur Verfügung.
Bei der analytischen Beurteilung standen für alle zwölf Variablen des Kriterienkatalogs in Anlehnung an Becker-Mrotzek und Böttcher jeweils die Ausprägungen 1, 0.5 und 0

zur Verfügung. Mit einem Punkt wurde die Variable bewertet, wenn das ausgewählte Kriterium vollkommen, mit 0.5 falls es in Ansätzen und mit 0 falls es überhaupt nicht vorhanden war. Die sich daraus ergebenen Punkte wurden anschließend zu einer Gesamtpunktzahl addiert, wobei maximal 13 Punkte zu erreichen waren. Für die Auswertung wurden darüber hinaus die Kategorien *Sprachrichtigkeit* und *Sprachangemessenheit* zur Oberkategorie *Sprache* zusammengefasst und der Aspekt *Prozess* zur Kategorie *Aufbau* gezählt, um einen Vergleich von analytischen und holistischen Verfahren zu ermöglichen.

Bei der holistischen Leistungsbeurteilung waren die Rater dazu angehalten, einerseits das gesamte Textprodukt und andererseits das Textprodukt im Hinblick auf die Kategorien *Sprache, Inhalt* und *Aufbau* zu benoten, sodass jedes Textprodukt insgesamt vier Noten erhielt. An dieser Stelle muss angemerkt werden, dass die Skalen für die einzelnen Teilbereiche eigentlich lediglich semi-holistisch sind, da nur einzelne Aspekte des Textes beurteilt werden. Im weiteren Verlauf werden wir dieses Vorgehen der Einfachheit halber allerdings ebenfalls als holistisch betrachten, da bei allen Beurteilungen der Text als Ganzes im Fokus steht. Jede der vier Noten sollte in einer einzelnen Runde vergeben werden, sodass sich insgesamt vier Benotungsrunden ergaben. Anschließend wurden die Noten mithilfe des Punktesystems für die gymnasiale Oberstufe in Punkte umgewandelt, um einen Vergleich mit dem analytischen Beurteilungsverfahren zu ermöglichen.

Alle Textprodukte wurden von drei Ratern beurteilt. Zwei Rater arbeiteten dabei holistisch, während ein Rater analytisch vorging und den entwickelten Kriterienkatalog nutzte. Da ist nicht unbedeutend ist, wer welches Verfahren der Leistungsbewertung nutzt, wurden die Aufgaben zugewiesen. Der Deutschlehrer der Klasse sowie die Referendarin, die zuvor in der Klasse unterrichtet hatte, arbeiteten mit holistischen Verfahren. Da ich die geringste Erfahrung mit Referenzkorpora hatte, arbeitete ich mit dem Kriterienkatalog.

3.3 Auswertung und statistisches Vorgehen

Die gesammelten Daten wurden anschließend sowohl mit der Statistiksoftware SPSS als auch mit Excel ausgewertet. Neben absoluten und relativen Häufigkeiten wurde dabei vor allem der ICC Koeffizient genutzt.

Die Intraklassenkorrelation (ICC) ist ein Verfahren zur Reliabilitätsbestimmung und kann genutzt werden, um die Übereinstimmung zwischen verschiedenen Ratern zu ermitteln. Voraussetzung für die Nutzung des dazugehörigen ICC Koeffizienten ist, dass intervallskalierte Werte vorliegen. Ein Wert von 0 impliziert, dass es keinen Zusammenhang zwischen den Ratern gibt. Je näher der Wert bei 1 liegt, desto höher ist der Zusammenhang zwischen den Ratern und damit die Reliabilität. Allgemein wird eine Intraklassenkorrelation als gut bewertet, wenn der Wert oberhalb von 0,7 liegt. Dies kann allerdings nur ein grober Richtwert sein, da Stichprobe und Merkmalsausprägungen mit einbezogen werden müssen (vgl. Wirtz & Caspar, 2002, S. 157f.).

4 Ergebnisse

Von den 23 möglichen Textprodukten konnten 22 ausgewertet werden. Eine Abgabe konnte aufgrund mangelnder Bearbeitung der Aufgabe nicht in die Analyse einbezogen werden.

Im Folgenden werden die Ergebnisse hinsichtlich der einzelnen Forschungsfragen nacheinander vorgestellt. Zunächst werden die Ergebnisse hinsichtlich der Interrater Übereinstimmung bei den holistischen Beurteilungsverfahren präsentiert und anschließend ein Vergleich zwischen holistischen und analytischen Verfahren gezogen.

4.1 Interrater Übereinstimmung bei holistischen Verfahren der Leistungsbeurteilung

Vergleicht man die vergebenen Noten der beiden Rater, die holistische Beurteilungsverfahren verwendet haben, so ist kaum zu übersehen, dass die Benotungen durchaus divergieren. Rater 2 benotet dieselben Textprodukte dabei sowohl bei der globalen Beurteilung des gesamten Textes als auch in den einzelnen Kategorien durchweg besser.

Tabelle **1:** Durchschnittliche Punktzahl der Rater bei holistischen Bewertung

	Sprache	Inhalt	Aufbau	Gesamt
Rater 1	10,23	10,09	9,82	9,68
Rater 2	10,55	10,96	10,86	10,46

So vergibt Rater 2 eine durchschnittliche Punktzahl von 10,46 Punkten bei der globalen Beurteilung des Textes, während Rater 1 die Textprodukte mit durchschnittlich 9,68 Punkten bewertet. Bemerkenswert ist dabei vor allem, dass beide Rater bei der globalen Beurteilung des gesamten Textprodukts die wenigsten Punkte vergeben und die einzelnen Kategorien durchschnittlich jeweils besser beurteilen.

Die Kategorien in denen die beiden Rater durchschnittlich die meisten Punkte vergeben unterscheiden sich allerdings. Rater 1 vergibt mit 10,23 Punkten die meisten Punkte in der Kategorie *Sprache*, während Rater 2 mit 10,96 Punkten die meisten Punkte im

Bereich *Inhalt* vergibt. Aufgrund dessen entstehen Differenzen zwischen den einzelnen Kategorien. Die geringste Differenz von 0,32 Punkten ist dabei im Teilbereich *Sprache* festzustellen, die größte Differenz von 1,04 Punkten in der Kategorie *Aufbau*.
Um die Reliabilität der Beurteilungen zu prüfen, lohnt es sich, den ICC Koeffizienten hinzu zu ziehen. Der Zusammenhang zwischen den beiden Ratern ist dabei sowohl bei der globalen Beurteilung des gesamten Textprodukts als auch innerhalb der einzelnen Kategorien sehr hoch, was auf eine hohe Reliabilität aller Beurteilungen schließen lässt. Der größte Zusammenhang ist mit einem Wert von 0,968 in der Kategorie *Sprache* festzustellen. In den Kategorien *Inhalt* und *Aufbau* ist der Zusammenhang am niedrigsten, wobei auch hier die Reliabilität noch ausgesprochen gut ist.

Tabelle **2:** Item-Inter-Korrelation bei holistischer Bewertung

Kategorie	ICC Koeffizienz
Sprache	0,968
Inhalt	0,926
Aufbau	0,934
Gesamt	0,946

Untersucht man die für die Textprodukte vergebenen Noten auf genaue Übereinstimmung, so konnte diese bei der globalen Benotung des gesamten Textproduktes nur in 22,7% aller Fälle festgestellt werden. Lässt man jedoch eine Abweichung von einem Notenpunkt nach oben oder unten zu, so ergibt sich eine Übereinstimmung von 81,8% zwischen beiden Ratern. Es lässt sich also festhalten, dass die vergebenen Noten bei der globalen Beurteilung des gesamten Textprodukts häufig nicht komplett übereinstimmen, aber mehrheitlich nur minimal voneinander abweichen.
Ähnliche Ergebnisse ergeben sich bei der Analyse der einzelnen Kategorien. Die höchste Übereinstimmung ergibt sich in der Kategorie *Sprache*, in der in 63,6% der Fälle die gleiche Note von beiden Rater vergeben wurde. Zieht man die Fälle mit einer Abweichung von einem Notenpunkt hinzu, entsteht sogar eine Übereinstimmung von über 90%. In den Kategorien *Aufbau* und *Inhalt* wurden hingegen nur 36,6% der Fälle gleich benotet.

Tabelle **3:** Übereinstimmung in Prozent bei holistischer Bewertung

	Sprache	Inhalt	Aufbau	Gesamt
Fälle mit komplette Übereinstimmung	63,6%	36,6%	36,6%	22,7%
Fälle mit kompletter Übereinstimmung & Abweichung von einem Notenpunkt	90,9%	63,6%	54,5%	81,8%

Es bleibt festzuhalten, dass sich Differenzen in der Beurteilung der beiden Rater feststellen lassen, wobei diese häufig minimal sind. In der Kategorie *Sprache* sind die Unterschiede am geringsten und die Reliabilität am größten. Im Folgenden sollen nun die Ergebnisse des Vergleichs zwischen analytischen und holistischen Beurteilungsverfahren präsentiert werden.

4.2 Vergleich von analytischen und holistischen Verfahren der Leistungsbeurteilung

Beim Vergleich des holistischen und des analytischen Beurteilungsverfahren muss auf relative Häufigkeiten zurückgegriffen werden, da sich die zu erreichenden Punkte in beiden Verfahren unterschieden.

Betrachtet man die globalen Bewertungen des gesamten Textprodukts, so schneiden die Schülerinnen und Schüler bei der analytischen Beurteilung mithilfe des Kriterienkatalogs am besten ab und erreichen durchschnittlich 71,9% der Punkte. Bei der holistischen Beurteilung des gesamten Textes durch die beiden Rater werden die Textprodukte hingegen mit durchschnittlich 64,5% bzw. 69,7% der Punkte bewertet.

Tabelle **4:** Durchschnittlich vergebene Punkte als relative Häufigkeit

	Sprache	Inhalt	Aufbau	Gesamt
Analytisch	0,691	0,727	0,741	0,719
Holistisch 1	0,682	0,673	0,655	0,645
Holistisch 2	0,703	0,730	0,724	0,697

Insgesamt liegen die Beurteilungen aller drei Rater relativ nah beieinander. Vor allem die holistische Beurteilung von Rater 2 und die analytische Beurteilung weisen lediglich minimale Differenzen auf. Allerdings beurteilt Rater 1 sowohl die einzelnen Kategorien als auch das gesamte Textprodukt jeweils mit den wenigsten Punkten und grenzt sich damit von den anderen Ratern ab.

Im Hinblick auf die einzelnen Kategorien wurden die Textprodukte der Lernenden teilweise von Rater 2 und teilweise bei der Verwendung des Kriterienkatalogs besser bewertet. Dabei zeigt sich ein ähnliches Ergebnis wie bereits beim Vergleich der beiden Rater, die mit holistischen Verfahren gearbeitet haben. Vor allem in der Kategorie *Aufbau* liegen die durchschnittlich vergebenen Punkte weiter auseinander, während sie in der Kategorie *Sprache* enger beieinander liegen.

In Abbildung 1 sind die globalen Bewertungen des gesamten Textprodukts in relativen Häufigkeiten als Grafik dargestellt. Auf der x-Achse sind die einzelnen Textprodukte aufgeführt, wobei sie aufsteigend nach der Punktzahl sortiert sind, die sie bei der analytischen Beurteilung erhalten haben. Auf der y-Achse ist die jeweils erreichte Punktzahl als relative Häufigkeit vermerkt.

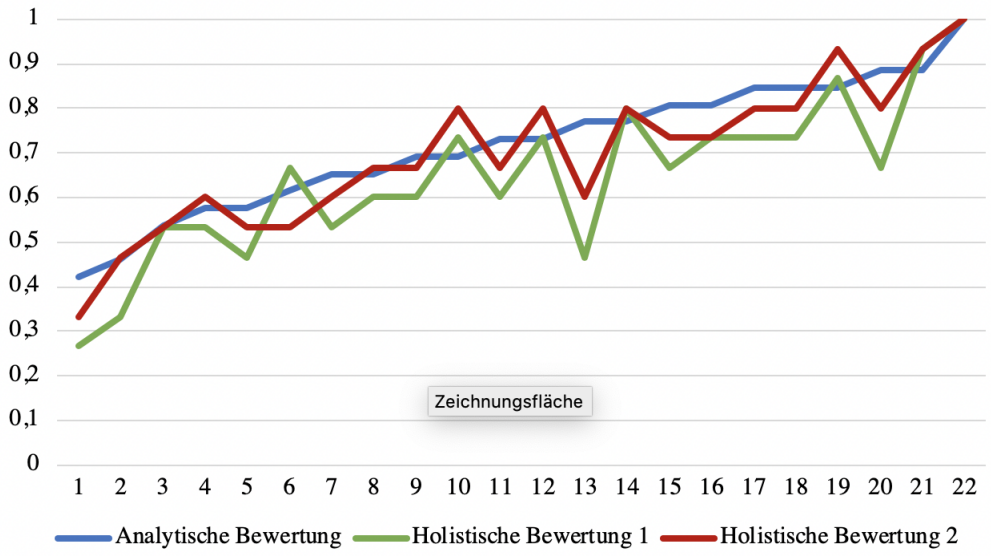


Abbildung **1:** Vergebene Punkte für das gesamte Textprodukt als relative Häufigkeit

Der engere Zusammenhang der analytischen Beurteilung und der holistischen Beurteilung von Rater 2 ist sichtbar. Allerdings fallen auch einige wenige Ausreißer auf, wobei das Textprodukt 13 am deutlichsten hervorsticht. Während dieses Produkt bei der analytischen Beurteilung 77% aller Punkte erhält, vergeben die Rater bei der holistischen Beurteilung lediglich 60% bzw. 47% aller verfügbaren Punkte. Bei genauer Betrachtung des Textprodukts wird sichtbar, dass das Textprodukt bei der Beurteilung mit dem Kriterienkatalog wenige Punkte in der Kategorie *Sprache* erhalten hat, aber in den Kategorien *Inhalt* und *Aufbau* überdurchschnittlich gut abgeschnitten hat. Aus diesem Grund ergibt sich bei der analytischen Bewertung insgesamt eine gute Note. Bei der holistischen Beurteilung wird das Textprodukt hingegen von beiden Ratern schlechter benotet.

Ein weiterer, nicht ganz so starker Ausreißer ist das Textprodukt 20. Auch hier erhält der Text bei der analytischen Beurteilung die meisten Punkte gefolgt von der Beurteilung von Rater 2. In diesem Beispiel erhielt allerdings die Kategorie *Inhalt* wenige Punkte in der analytischen Beurteilung, während die anderen Kategorien beinahe die volle Punktzahl erhalten haben.

Zieht man zur Bewertung der Reliabilität den ICC Koeffizienten hinzu, so fällt auf, dass der Zusammenhang zwischen den beiden Ratern, die mit holistischen Verfahren gearbeitet haben, mit einem Wert von 0,946 am größten ist. Die analytische Beurteilung und die Beurteilung von Rater 1 weisen mit einem Wert von 0,857 hingegen den geringsten Zusammenhang auf. Trotzdem ist dieser Wert noch ausreichend groß, um von einer guten Reliabilität sprechen zu können.

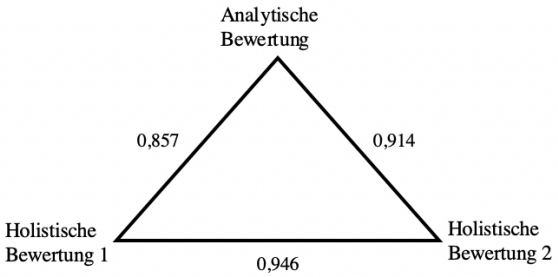


Abbildung **2:** Item-Inter-Korrelation zwischen analytischer und holistischer Bewertung

Vergleicht man die Rangordnung, die durch die unterschiedlichen Verfahren der Leistungsbewertung in die Textprodukte der Schülerinnen und Schüler gebracht wird, so lassen sich auch hier Unterschiede ausmachen. Unabhängig von der Art der Beurteilung wurden die selben zwei Textprodukte als die beiden besten und zwei weitere als die beiden schwächsten Produkten bewertet. Abweichungen gibt es vor allem im Mittelfeld, sodass je nach Rater und Art der Leistungsbeurteilung eine unterschiedliche Rangordnung innerhalb der Klasse zustande kommt. Häufig sind diese Unterschiede bis auf einige Ausreißer allerdings minimal.

Tabelle **5:** Rangfolge innerhalb der Klasse in aufsteigender Punktzahl

Analytisch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Holistisch 1	1	2	5	13	3	4	7	8	9	11	6	15	20	10	12	16
Holistisch 2	1	2	3	5	6	4	7	13	8	9	11	15	16	10	12	14

Analytisch	17	18	19	20	21	22
Holistisch 1	17	18	14	19	21	22
Holistisch 2	17	18	20	19	21	22

Die hier dargestellten Ergebnisse sollen anschließend im folgenden Kapitel diskutiert und auf die theoretischen Grundlagen aus Kapitel 2 bezogen werden.

5 Diskussion

Insgesamt wurden die Textprodukte der Schülerinnen und Schüler sowohl mithilfe analytischer als auch mithilfe holistischer Verfahren mehrheitlich gut beurteilt. Dies lässt sich möglicherweise damit begründen, dass die Klassenarbeit mit einer vergleichbaren Fragestellung bereits im Vorfeld geschrieben worden war. Ein anderes Argument könnte allerdings auch sein, dass der Kriterienkatalog für die analytische Beurteilung zuvor gemeinsam mit den Lernenden ausgearbeitet worden ist, sodass die Kriterien den Schüler*innen sehr klar waren und sie explizit auf diese geachtet haben. Dieses Ergebnis würde dafür sprechen, auch im schulischen Alltag häufiger die Kriterien explizit mit den Schüler*innen zu erarbeiten oder zu thematisieren, um die Transparenz der Beurteilung zu erhöhen und die Lernenden optimal auf die Prüfung vorzubereiten.
Vergleicht man die Beurteilungen der beiden Rater, die holistische Verfahren der Leistungsbeurteilung verwendet haben, so kann die Hypothese bestätigt werden, dass sich die Benotungen der beiden Rater unterscheiden. Allerdings ist an dieser Stelle erst eine Analyse der einzelnen Kategorien aufschlussreich. So weist die Kategorie *Sprache* sowohl die geringste Differenz in der durchschnittlich vergebenen Punktzahl als auch die mit Abstand größte prozentuale Übereinstimmung auf.
Diese Ergebnisse lassen die Vermutung zu, dass in der Kategorie *Sprache* ähnliche Maßstäbe angelegt werden und es deswegen leichter fällt, die Textprodukte mit gleichen oder nur minimal abweichenden Noten zu bewerten. Ein Grund für diesen Umstand kann die Tatsache sein, dass bei der Kategorie *Sprache* leicht auf oberflächlich sichtbare Kriterien wie die Rechtschreibung zurückgegriffen werden kann. Dass eine reine Betrachtung orthografischer Fehler eine Verkürzung der Kategorie *Sprache* zur Folge hat, soll an dieser Stelle nicht weiter diskutiert werden. Trotzdem bleibt festzuhalten, dass die Kategorie *Sprache* anscheinend Eigenschaften aufweist, welche eine einheitliche Beurteilung erleichtert. Im Gegensatz dazu scheinen die Kategorien *Inhalt* und*Aufbau* weniger solcher allgemeingültiger Kriterien aufzuweisen, sodass die Rater häufiger auf individuelle Maßstäbe zurückgreifen müssen. Dementsprechend weisen die Beurteilungen dieser Kategorien größere Differenzen auf, da diese individuellen Maßstäbe keinesfalls gleich sein müssen.

Dieses Ergebnis geht ebenfalls aus der Betrachtung des ICC Koeffizienten hervor, da die Kategorie *Sprache* die höchste Reliabilität aufweist. In allen drei Kategorien zeigt sich generell ein überdurchschnittlich guter Zusammenhang zwischen den Ratern. Allerdings muss an dieser Stelle darauf hingewiesen werden, dass dies nicht bedeutet, dass die Noten gleich oder nah beieinander sind. Der ICC Koeffizient zeigt lediglich, dass die Beurteilungen einen beständigen Zusammenhang aufweisen. Folglich bedeutet dies aber auch, dass beide Rater die gesamte Klasse an den gleichen Maßstäben gemessen haben, was für eine faire Benotung nicht zu vernachlässigen ist.

Hinsichtlich des Vergleichs analytischer und holistischer Verfahren der Leistungsbeurteilung zeigt sich, dass die Lernenden bei der Beurteilung mit dem Kriterienkatalog durchschnittlich besser abschneiden als bei den holistischen Beurteilungen des gesamten Textes. Ein Grund für dieses Ergebnis könnte in der Tatsache liegen, dass sich die beiden Verfahren grundsätzlich unterscheiden. Bei holistischen Verfahren wird das Textprodukt als Ganzes verstanden und nach Fehlern durchsucht, die dieses Textprodukt von einem sehr guten Textprodukt unterscheiden. Die analytische Leistungsbeurteilung geht anders vor, indem sie jedes Kriterium einzeln betrachtet und vermerkt, ob dieses vorhanden ist oder nicht. Dieser Unterschied könnte dazu führen, dass bei einer holistischen Beurteilung die Fehler stärker ins Gewicht fallen, während bei einer analytischen Beurteilung eher die erreichten Variablen im Fokus stehen. Darüber hinaus wertschätzt der Kriterienkatalog einzelne Variablen, die bei einer holistischen Beurteilung möglicherweise vernachlässigt werden. So kann beispielsweise beim Kriterienkatalog ein eigener Punkt für das Wählen einer passenden Überschrift vergeben werden, während diese Tatsache bei der holistischen Beurteilung möglicherweise als selbstverständlich erachtet wird. Ob diese Vermutungen zutreffen, gilt es weiter zu untersuchen, da sie im Rahmen dieser Arbeit weder bestätigt noch widerlegt werden können.

Trotz aller Unterschiede zeigen die Beurteilungen der Rater aber doch einen großen Zusammenhang. Dieser ist zwischen den beiden Ratern, die mit holistischen Verfahren gearbeitet haben am größten. Dies ist ein Indiz dafür, dass bei holistischen Beurteilungen ähnliche Maßstäbe vorzuliegen scheinen, die sich anscheinend von analytischen Maßstäben in Ansätzen unterscheiden. Darüber hinaus zeigen die analytische Beur-

teilung und die holistische Beurteilung von Rater 2 einen sehr guten Zusammenhang auf. Diese Beurteilungen stammen von dem Lehrer der Klasse, der zwei Vorteile gegenüber den anderen beiden Ratern hat. Einerseits kennt er die Klasse am besten und andererseits hat er durch jahrelange Arbeit in der Schule die meiste Erfahrung mit Referenzkorpora.

Die Hypothese, dass die Bewertungen vor allem im Mittelfeld differieren, kann für diese Gruppe bestätigt werden. Während alle Rater dieselben Textprodukte als die beiden besten und die beiden schwächsten identifizieren, unterscheidet sich die Rangfolge der Textprodukte im Mittelfeld von Rater zu Rater. Wie in der Forschung bereits gezeigt, sind in diesem Fall vor allem einzelne Kategorien ausschlaggebend, da die Rater häufig leicht zu bewertende Kriterien heranziehen. Darüber hinaus zeigen die Ergebnisse, dass es wohl besonders anspruchsvoll ist, Textprodukte zu benoten, die in einzelnen Kategorien Defizite aufweisen, aber in anderen Kategorien mindestens gut sind. Dieser Umstand lässt sich anhand der Ausreißer beobachten, da die Beurteilungen dieser Textprodukte am stärksten differieren. Die Rater, die mit holistischen Verfahren gearbeitet haben, beurteilten die Textprodukte dabei häufig mit weniger Punkten als der Rater, der mit Kriterienkatalog gearbeitet hat. Dies lässt die Vermutung zu, dass deutliche Defizite in einem Teilbereich einen starken Einfluss auf die Gesamtwahrnehmung des Textes haben und diese negativ beeinflussen. Allerdings ist bemerkenswert, dass nicht allein die Kategorie *Sprache* ausschlaggebend ist, sondern auch Defizite in anderen Kategorien zu diesem Ergebnis führen. Zeigen Lernende in einem Teilbereich also deutlich schwächere Leistungen als in den anderen Bereichen, so fällt die schwache Leistung besonders stark ins Gewicht.

6 Fazit

Abschließend lässt sich festhalten, dass die Ergebnisse die formulierten Forschungshypothesen größtenteils bestätigen. In diesem Zusammenhang muss natürlich beachtet werden, dass die Stichprobe verhältnismäßig klein ist und die Ergebnisse aus diesem Grund keineswegs repräsentativ sind. Nichtsdestotrotz lassen sich interessante Schlussfolgerungen formulieren, die eine Grundlage für weitere Forschungsfragen sein und zur Reflexion über die eigene Beurteilungspraxis anregen können.

Zunächst konnte in dieser Untersuchung festgestellt werden, dass die Unterschiede zwischen holistischen und analytischen Verfahren nicht so auffällig sind, wie man es zunächst möglicherweise vermuten würde. Vor allem die holistische Beurteilung von erfahrenen Lehrkräften und die analytische Beurteilung weisen deutliche Ähnlichkeiten auf. Gab es gravierende Differenzen zwischen den vergebenen Noten, so konnte diese Tatsache mehrfach auf Schwächen in einem Teilbereich zurückgeführt werden. Deutliche Schwächen in einer Kategorie können demzufolge vor allem bei holistischen Verfahren die Gesamtnote negativ beeinflussen. An dieser Stelle wäre eine Untersuchung interessant, die den Einfluss der einzelnen Kategorien auf die Gesamtnote misst, aber in dem Kontext dieses Forschungsprojekts zu weit führen würde. Spannend ist des Weiteren vor allem der Befund, dass die Kategorie *Sprache* anscheinend leichter zu bewerten ist als andere Kategorien und die Bewertungen hier am häufigsten übereinstimmen.

Darüber hinaus ließen sich durch einige Veränderungen ausschlaggebendere Ergebnisse gewinnen. So war allen Ratern die Klasse vorab bekannt, sodass trotz des Verzichts auf Namen davon ausgegangen werden muss, dass den Ratern die Handschrift der Schüler*innen bekannt war. Darüber hinaus wurden Variablen wie Handschrift und Länge eines Textprodukts in der Analyse der Ergebnisse außen vor gelassen. Insbesondere eine umfangreichere Auswahl an Textprodukten sowie Ratern könnte repräsentativere Ergebnissen herbeiführen. Durch zusätzliche Rater wäre es möglich, auch innerhalb des analytischen Beurteilungsverfahrens auf Interrater Reliabilität zu prüfen. Durch mehrere Textprodukte der gleichen Lernenden könnte darüber hinaus die Intrarater Reliabilität untersucht werden, also ob Lehrkräfte in der Beurteilung einzelner Schüler*innen konsistent sind.

Literaturverzeichnis

Bacha, N. (2001): *Writing evaluation: what can analytic versus holistic essay scoring tell us?* In: System, **29**, 3: S. 371 – 383.

Baurmann, J. (2006): *Schreiben - Überarbeiten - Beurteilen. Ein Arbeitsbuch zur Schreibdidaktik.* 2. Aufl. Kallmeyer Verlag in Verbindung mit Klett, Seelze.

Becker-Mrotzek, M. (2008): *Texte beurteilen - transparent und gerecht.* In: Schulverwaltung. Nordrhein-Westfalen, **19**, 6: S. 175–176.

Becker-Mrotzek, M. & Böttcher, I. (2009): *Texte bearbeiten, bewerten und beurteilen.* 4. Aufl. Cornelsen-Verlag, Berlin.

Becker-Mrotzek, M. & Böttcher, I. (2015): *Schreibkompetenz entwickeln und beurteilen.* 6. Aufl. Cornelsen-Verlag, Berlin.

Birkel, C. & Birkel, P. (2002): *Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss.* In: Psychologie in Erziehung und Unterricht, **49**, 3: S. 219–224.

Dehn, M. & Baurmann, J. (2004): *Beurteilen im Deutschunterricht.* In: Praxis Deutsch, **31**, 184: S. 6–13.

Grausam, N. C. (2018): *Diagnosekompetenz von Lehrpersonen als Voraussetzung individueller Förderung im Bereich „Texte schreiben". Eine empirische Studie am Beispiel einer neu eingeführten integrierten Schulform.* Waxmann-Verlag, Münster.

Nussbaumer, M. & Sieber, P. (1994): *Texte analysieren mit dem Zürcher Textanalyseraster.* In: Sieber, P. [Hrsg.]: *Sprachfähigkeiten - besser als ihr Ruf und nötiger denn je! Ergebnisse und Folgerungen aus einem Forschungsprojekt*, S. 141–186. Verlag Sauerländer, Aarau.

Schipolowski, S. & Böhme, K. (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.

Wirtz, M. & Caspar, F. (2002): *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen.* Hogrefe-Verlag, Göttingen.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang

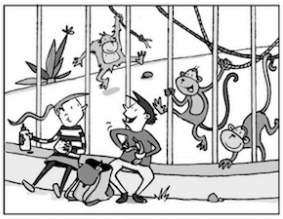
Cremer | Eine Erzählung planen und gestalten| 5D 03.07.2019

Eine Bildergeschichte verfassen

Aufgabe 1

Verfasse eine Bildergeschichte zu den abgebildeten Bildern.

a) Bringe die Bilder zunächst in die richtige Reihenfolge.



b) Plane deine Erzählung. Lege dazu ein Ideennetz oder eine Stichwortliste auf diesem Arbeitsblatt an.

c) Schreibe die Geschichte und finde eine passende Überschrift. Die Geschichte sollte mindestens eine Seite lang sein. Du darfst gerne kreativ sein und die Bildergeschichte mit eigenen Ideen ergänzen.

Anhang

Benotung Bildergeschichte

Benotung Kriterienkatalog:

Dimension	Kriterium	1	0,5	0
Sprach-richtigkeit	Du verwendest bekannte Rechtschreibregeln richtig.			
	Du bildest die Sätze richtig und verständlich.			
Sprachan-gemessenheit	Du verwendest wörtliche Rede.			
	Du verwendest eine einheitliche Zeitform.			
	Du beschreibst, was die Figuren denken und fühlen.			
Inhalt	Du wählst eine passende, neugierig machende Überschrift.			
	Du erzählst eine vollständige, zusammenhängende, interessante Geschichte.			
	Du beachtest die Handlung und die Reihenfolge der Bilder.			
Aufbau	Du beantwortest in der Einleitung die Fragen Wer, Wann und Wo.			
	Du beschreibst im Hauptteil den Höhepunkt der Geschichte.			
	Du formulierst einen passenden Schluss.			
	Du beschreibst die Handlung zwischen den Bildern.			
Prozess	Du planst und überarbeitest deinen Text erkennbar.			

Benotung Rater 1 und Rater 2:

	Rater 1	Rater 2
Note Sprache		
Note Inhalt		
Note Aufbau		
Gesamtnote		