Bibliographischer Hinweis sowie Verlagsrechte bei den online-Versionen der DD-Beiträge:

Didaktik Deutsch

Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur
http://www.didaktik-deutsch.de
8. Jahrgang 2003 – ISSN 1431-4355
Schneider Verlag Hohengehren GmbH

Peter Birkel

AUFSATZBEURTEILUNG – EIN ALTES PROBLEM NEU UNTERSUCHT

In: Didaktik Deutsch. Jg. 8. H. 15. S. 46-63.

Peter Birkel

AUFSATZBEURTEILUNG – EIN ALTES PROBLEM NEU UNTERSUCHT

1 Begründung der Replikationsstudie

Wenn man bei Vorträgen Lehrerinnen und Lehrern der verschiedenen Schularten etwas über die Problematik der Notengebung berichtet und sich dabei auf die Literatur zu diesem Themenbereich stützt, dann sieht man sich u.U. schnell dem Vorwurf ausgesetzt, dass die Beispiele, die man für die mangelnde Reliabilität und Validität der Ziffernbenotung anführt, inzwischen ja ziemlich alt seien. Ingenkamp hat sein Buch über „die Fragwürdigkeit der Zensurengebung" in der ersten Auflage bereits 1971 herausgegeben. Die darin publizierten Originalarbeiten sind noch viel älter und stammen z.T. aus den anglo-amerikanischen Ländern. Immer wieder werden Zweifel gehegt, ob die dort berichteten Ergebnisse noch Relevanz für die heutige Schulsituation in Deutschland haben können. Wurde in den vergangenen dreißig bis vierzig Jahren die Validität solcher Untersuchungen nicht einfach dadurch reduziert, dass immer mehr Lehrer bereits während ihrer eigenen Ausbildung ausführlich über diese Problematik informiert wurden? Haben denn nicht in der jüngeren Vergangenheit immer mehr Lehrer Möglichkeiten zur Objektivierung der Leistungsbeurteilung kennen gelernt? Nur zu oft wird einem vorgehalten, dass die berichteten Probleme die heutige Beurteilungspraxis gar nicht mehr beträfen!

Wie sieht es aber um die Ausbildung der Lehrer im Bereich Leistungsmessung und Leistungsbeurteilung aus? Was lernen Lehramtsanwärter heute wirklich an den Hochschulen, um ihre Notengebung auf eine zuverlässigere Basis zu stellen? Aus den Erfahrungen an einer Pädagogischen Hochschule, die Lehrer für die Grund- und Hauptschule und für die Realschule ausbildet, kann berichtet werden, dass in den Fächern Psychologie und Pädagogik zwar immer wieder Lehrveranstaltungen zum Problembereich Notengebung angeboten werden, dass aber aufgrund der Wahlfreiheit bei der Zusammenstellung von Stundenplänen durch die Studenten solche Lehrveranstaltungen nur von einem Bruchteil der Lehramtsanwärter tatsächlich besucht werden.[1] Fachdidaktische Lehrveranstaltungen im Fach Deutsch greifen diese Thematik eher randständig auf. Die Lehramtsstudenten ebenso wie die bereits tätigen Lehrer glauben, dass man die Notengebung nicht besonders erlernen muss, weil die-

[1] Neben vielen objektiv vorhandenen Schwierigkeiten bei der Stundenplanzusammenstellung kommt vor allen Dingen ein wesentliches Problem hinzu: Wer sich über die Problematik der Ziffernbenotung informieren will, muss in der Lage sein, empirische Untersuchungen mit all ihren statistischen Angaben zu lesen. Eine Einführung in die Statistik machen Lehramtsstudenten aber nur in seltenen Ausnahmefällen mit. Wer schlägt sich – vor allem als Studierender des Faches Deutsch - schon gern mit mathematischen Formeln herum, wenn es so viele interessante Lehrangebote im eigenen Studienfach und im erziehungswissenschaftlichen Teil des Studiums gibt? Gegen Thomas Manns „Zauberberg" oder die Freudsche Theorie der sexuellen Entwicklung im Kindesalter kommt eine „trockene" Statistikveranstaltung kaum an.

se doch ganz einfach nur so fortgeführt werden müsse, wie man sie schon früher als Schüler bei seinen eigenen Lehrerinnen und Lehrern erlebt hat. Diese Erfahrungen stützen die Hypothese, dass so alte Forschungsergebnisse, wie z.B. die zur Problematik der Aufsatzbeurteilung, vermutlich auch heute noch ihre Gültigkeit haben.[2]

Die Beurteilung von Aufsätzen wurde in den Jahren zwischen 1965 und heute nicht oft untersucht. Wenn es zu Untersuchungen kam, dann wurden vor allem die von Ingenkamp (1971) geforderten und von Beck (1974, 1975) vorgeschlagenen Kriterien zur Aufsatzbeurteilung berücksichtigt. In deren Folge wurden immer wieder Überlegungen angestellt, wie man zu Bewertungskriterien gelangen könnte, die u.U. nicht nur aufgrund fachdidaktischer Überlegungen begründbar, sondern in ihrem Wert sogar empirisch belegbar wären (vgl. Lehmann 1990, S. 71f). Auf jeden Fall belegen die Untersuchungen zur Aufsatzbeurteilung, die die Verwendung solcher Kriterienkataloge benutzten, dass damit die Reliabilität bei Zensuren so sehr gesteigert werden kann, dass sie nahe an die Größenordnung heran kommt, die für formelle Tests gefordert wird (vgl. Beck 1974, Beck/Hofen 1991, Fliegner 1974, Grzesik/Fischer 1984, Lehmann 1987, 1988, 1990a, 1990b, 1993, 1994a, 1994b, Nussbaumer 1991, Weber 1973). Solche Untersuchungen wurden aber vornehmlich im Bereich der Sekundarstufe durchgeführt. Eigene Erfahrungen mit der Aufsatzbeurteilung in der Grundschule machen aber deutlich, dass die Anwendung solcher Kriterienkataloge hier nur in seltenen Ausnahmenfällen Anwendung findet. Deshalb wird auch bei dieser Untersuchung auf die Vorgabe solcher Kriterien verzichtet.

2 Untersuchungsmaterial

Im Rahmen des Tagespraktikums der Lehramtsstudenten an der PH Weingarten suchten die Studierenden aus einem Aufsatz, der zufällig kurz vor Beginn des Praktikums von den Schülern der vierten Grundschulklasse geschrieben worden war, vier Aufsätze heraus, die man im Rahmen der Untersuchung verwenden wollte. Aufgabe der Schüler war es gewesen, einen „Reizwort-Aufsatz" zu schreiben, in den die drei Worte *„Langeweile - Dachboden - Kleidertruhe"* zu integrieren waren. Auch eine Überschrift dazu sollte gefunden werden.

Die Aufsätze sollten sich hinsichtlich der Qualität deutlich unterscheiden, was an den ursprünglich gegebenen Noten abgelesen wurde. Zusätzlich wurde darauf geachtet, dass sich die Aufsätze auch hinsichtlich der Länge unterschieden. Es wurden ein längerer (Aufsatz 1 = 385 Wörter), zwei mittellange (Aufsatz 2 = 171 Wörter,

---

2 Im Rahmen einer Lehrveranstaltung an der PH Weingarten wurde das von den Studierenden heftig angezweifelt. Daraus erwuchs dann der Wunsch, eine solche Untersuchung in der heutigen Zeit einmal zu wiederholen. Aus der Sorge heraus, dass man bei einer Nachuntersuchung in einem anderen (objektiver messbaren?) Fach wahrscheinlich zu keinem Ergebnis käme, einigte man sich auf die Aufsatzbeurteilung, weil diese nach der Meinung der Studenten eher subjektiven Ermessensspielraum eröffnete. An die Replizierbarkeit der Ergebnisse mit einer Rechenarbeit (wie bei Weiss 1966b) glaubten die Studenten von vornherein nicht.

Aufsatz 3 = 238 Wörter) und ein kurzer Aufsatz (Aufsatz 4 = 95 Wörter) ausgesucht.[3]

Zusätzlich sollte untersucht werden, inwieweit die Beurteilung des Aufsatzes abhing von der Rechtschreibfähigkeit des betreffenden Schülers. Dazu wurden die Originaltexte der Schüler je zweimal in maschinengeschriebene Form gebracht, eine Version mit relativ wenig Fehlern, wenngleich auch nicht völlig fehlerfrei, und eine Version mit vielen Fehlern. Hier wurden alle im Original vorhandenen Fehler beibehalten und weitere typische Fehler für diese Altersstufe eingefügt. Gleichzeitig beurteilte jeder Lehrer einen langen Aufsatz (1), zwei mittellange Aufsätze (2+3) und einen kurzen Aufsatz (4).

3 Untersuchungsansatz

Jeder Lehrer, der an der Untersuchung teilnahm, sollte alle vier Aufsätze beurteilen. Um nun eine einigermaßen systematische Variation der Variable Fehlerhaftigkeit zu erreichen, wurden die Aufsätze in acht verschiedenen Konstellationen angeordnet. Jede Konstellation sollte von einer Lehrergruppe beurteilt werden. In Konstellation 1 waren alle vier Aufsätze mit wenig Fehlern zu beurteilen, in den Konstellationen 2 bis 7 enthielten jeweils zwei Aufsätze wenig und zwei viele Fehler, und in Konstellation 8 waren alle vier Aufsätze mit vielen Fehlern zu beurteilen (s. Tab.1 unten). Durch diese Anordnung sollte gewährleistet werden, dass bei in etwa gleichen Gruppengrößen annähernd gleich viele Lehrer die Aufsätze mit vielen oder mit wenigen Fehlern zu beurteilen hatten.

Gruppe	1	2	3	4	5	6	7	8	
Aufs. 1	wenig[4]	wenig	wenig	wenig	viel	viel	viel	viel	
Aufs. 2	wenig	wenig	viel	viel	viel	wenig	wenig	viel	
Aufs. 3	wenig	viel	wenig	viel	wenig	viel	wenig	viel	
Aufs. 4	wenig	viel	viel	wenig	wenig	wenig	viel	viel	
N =	7	9	14	8	10	8	12	21	Σ=89

Tab. 1: Design der Untersuchung

4 Stichprobe

Die eigentliche Untersuchung fand im weiteren Umfeld der PH Weingarten zwischen Bodensee und Donau statt. Insgesamt waren die Beurteilungen von 89 Lehrerinnen und Lehrern zusammen gekommen. Sie verteilten sich auf die acht ver-

---

3 Die Originalaufsätze sind im Anhang abgedruckt.

4 viel = viele Fehler, wenig = wenig Fehler

schiedenen Beurteilungskonstellationen wie aus Tab. 1 zu ersehen ist. Gruppe 8, die alle vier Aufsätze mit vielen Fehlern zu beurteilen hatte, stellte mit fast einem Viertel aller Lehrer die größte Gruppe. Darum lag die Gruppegröße der Lehrer, die die Aufsätze mit vielen Fehlern zu beurteilen hatte deutlich über der der Lehrer, deren Aufsätze nur wenig Fehler enthielten.

Auf eine Aufgliederung nach Geschlecht und Alter der beteiligten Lehrer wurde verzichtet, da den beurteilenden Lehrern absolute Anonymität zugesichert war. Es dürfte sich aber zum größten Teil um Lehrerinnen handeln, die bis auf wenige Ausnahmen zum Zeitpunkt der Untersuchung eine vierte Grundschulklasse unterrichteten.

5 Ergebnisse

Unterschiede zwischen Lehrergruppen

In einem ersten Anlauf wurden die Beurteilungsunterschiede zwischen den acht Lehrergruppen und den vier Aufsätzen in einer zweifaktoriellen Varianzanalyse mit Messwiederholung auf einem Faktor auf Signifikanz überprüft. Die Messwiederholung lag darin, dass ja jeder Lehrer alle vier Aufsätze zu beurteilen hatte. Das Ergebnis ist in Tab. 2 zusammengefasst.

Quelle	SAQ	df	Varianz	F	p <	Effekt%
Zw. Subj.	714032	88	8114,0			
Zw. (Lehrer)Gruppen	23040	7	3291,4	0,386	n.s.	0,5
In. (Grup.)	690992	81	8530,8			
In. Subj.	3719680	267	13931,4			
Zw. Messwied. (Aufsätze)	2977331	3	992443,7	407,652	.001	80,0
Zw. Gruppe*Messwied.	150757	21	7178,9	2,949	.001	4,1
In. (Messwied.)	591592	243	2434,5			
Total	4433712	355	12489,3			

Tab. 2: Varianzanalyseergebnisse[5]

Zunächst kann man sich anschauen, ob sich die Lehrergruppen, die ihre Beurteilungen abgegeben haben, generell in ihrem Beurteilungsverhalten unterschieden haben. Der Unterschied zwischen den Gruppen ist mit $F=0.386$ fern jeglicher Signifikanz. Gemittelt über alle vier Aufsätze unterscheiden sich die beteiligten Lehrergruppen *nicht* in ihrer Einschätzung der Qualität der Aufsätze.

---

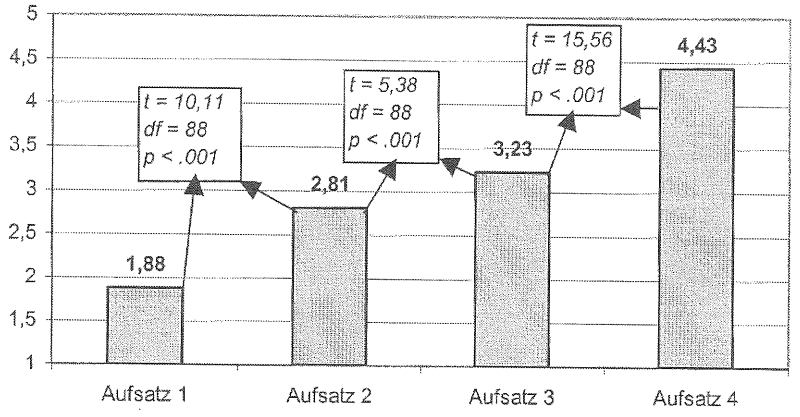
5 Entscheidend für die Signifikanz eines Effekts ist der Wert F. In der nächsten Spalte wird das Signifikanzniveau angegeben. In der letzten Spalte wird angegeben, wieviel % der Gesamtvarianz durch die jeweilige Quelle der Varianz geklärt wird.

Qualität der Aufsätze

Als nächstes kann festgestellt werden, dass die verwendeten Aufsätze tatsächlich signifikant unterschiedlicher Qualität waren (*F=407,7; df=3/243; p<.001*). Die Qualität der Aufsätze klärt allein 80% der Gesamtvarianz. Man darf somit davon ausgehen, dass sich die Lehrer in ihrem Urteil tatsächlich in erster Linie an der unterschiedlichen Qualität der Aufsätze orientierten. Die durchschnittlichen Beurteilungen der Aufsätze waren wie in Abb. 1 dargestellt ausgefallen:

Man kann festhalten, dass der erste Aufsatz qualitativ der beste war. Seine Durchschnittsbeurteilung lag bei der Note 1,88, was praktisch einer 2+ entspricht. Das war auch die von der Klassenlehrerin ursprünglich erteilte Note. Die Aufsätze 2 und 3 liegen eher im mittleren Notenbereich. Dabei entspricht wiederum die 2,81 für den zweiten Aufsatz der Note 3+ und die 3,23 für den dritten Aufsatz der Note 3-. Auch diese Notenmittel entsprechen den ursprünglich erteilten Noten. Der vierte Aufsatz bekam mit dem Mittelwert 4,43 im Schnitt eine 4-5. Hier hatte die Klassenlehrerin eine 4- erteilt. Die Notenmittel der Lehrergruppen stimmen also erstaunlich gut mit den von der Klassenlehrerin ursprünglich erteilten Noten überein.

Abb. 1: Durchschnittsnoten der vier Aufsätze

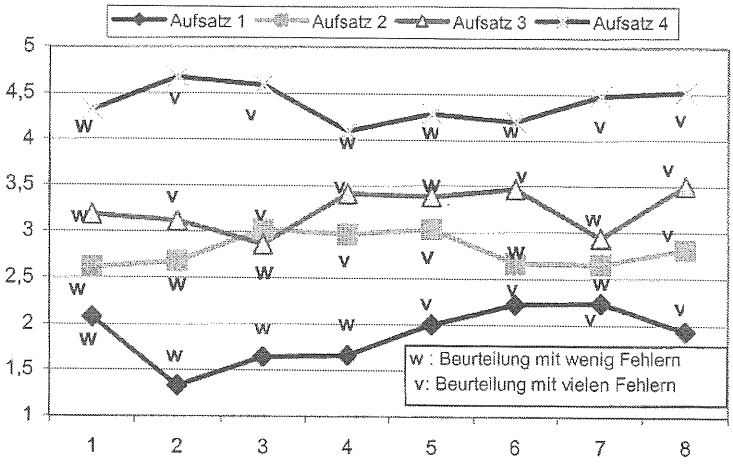


Eine Überprüfung der einzelnen Mittelwertsunterschiede auf Signifikanz mit dem t-Test für abhängige Stichproben ergab, dass sich alle Mittelwertsunterschiede auf dem 1‰-Niveau signifikant unterscheiden. Abb. 1 enthält nur die Ergebnisse des t-Tests für benachbarte Mittelwerte, wegen der größeren Mittelwertsdifferenzen bei allen weiteren Mittelwertsvergleichen steigt bei allen anderen t-Tests die Signifikanz noch weiter an.

Schließlich wird die Interaktion Gruppe*Aufsatz signifikant (s. Tab. 2). Hier wird man also darauf achten müssen, ob die Qualitätsabstufungen der Aufsätze in allen Lehrergruppen in gleicher oder ähnlicher Weise zum Tragen kommen. Die Mittelwerte fanden Aufnahme in Abb. 2 auf der nächsten Seite.

Um deutlicher die Beurteilungsbedingungen der Lehrer in den acht verschiedenen Gruppen hervortreten zu lassen, wurde jeweils der Aufsatz mit wenigen Fehlern mit einem (w) markiert, wenn er mit wenig Fehlern, und mit einem (v), wenn er mit vielen Fehlern zu beurteilen war. Bis auf eine Gruppe beurteilten die übrigen Lehrer die Aufsätze 1 bis 4 mit jeweils absteigender Qualität. Nur bei Gruppe 3 kam es vor, dass die Aufsätze 2 und 3 in „verkehrter" Reihenfolge qualitativ eingeordnet wurden. Hier wurde Aufsatz 2 mit vielen Fehlern als etwas schlechter eingestuft als Aufsatz 3, der in diesem Fall mit wenig Fehlern zu beurteilen war und hier die positivste Durchschnittsbeurteilung insgesamt erhielt. Ansonsten ist der Abb. 2 zu entnehmen, dass die Spannweite der Durchschnittsbeurteilungen relativ deutlich variiert mit der Qualität der Rechtschreibleistung. Fast immer verbessert sich die durchschnittliche Beurteilung, wenn ein Aufsatz mit wenig Fehlern zu beurteilen war, und sie verschlechtert sich, wenn der jeweilige Aufsatz viele Fehler enthielt. Eine Ausnahme von dieser Regel ist nur feststellbar, wenn die Lehrer alle Aufsätze mit wenigen (Gruppe 1) oder alle mit vielen Fehlern (Gruppe 8) zu beurteilen hatten.

Abb. 2 : Durchschnittsnoten für die vier Ausätze in den acht Gruppen



Bedeutsam erscheint auch, dass die Lehrer, die nur Aufsätze mit wenig Fehlern (Gruppe 1), von daher also eigentlich nur die Qualität der Textproduktion zu beurteilen hatten, Urteile abgaben, die relativ eng beieinander lagen. Lehrer, die nur Aufsätze mit vielen Fehlern zu beurteilen hatten (Gruppe 8), trafen dagegen am besten die Durchschnittsbeurteilungen aller Lehrer für die vier Aufsätze. Die größten Beurteilungsunterschiede ergaben sich dann, wenn der beste Aufsatz mit wenig Fehlern und der schlechteste Aufsatz mit vielen Fehlern beurteilt wurde. Das war in den Gruppen 2 und 3 der Fall. Andererseits lagen die Beurteilungen am engsten beieinander, wenn der beste Aufsatz viele Fehler und der schlechteste wenig Fehler enthielt (Gruppe 6).

Rechtschreibleistung, Aufsatzlänge und Note

Um bei dem relativ verschachtelten Design den Einfluss des Faktors „Rechtschreibleistung" auf die Notengebung leichter abschätzen zu können, wurde eine Varianzanalyse gerechnet mit den beiden Faktoren A „Länge des Aufsatzes" (3 Abstufungen: lang – mittel – kurz) und B „Fehlerhaftigkeit des Textes" (2 Abstufungen: wenig – viel) berechnet. Da die Messwiederholung bei der Aufsatzbeurteilung hier unberücksichtigt blieb, sind die Signifikanzen, sofern sie auftreten, mit größerer Sicherheit wirklich signifikant, da wir bei der Annahme der Unabhängigkeit der Gruppen eher gegen uns gearbeitet haben. Das Ergebnis der Varianzanalyse ist in Tab. 3 auf der nächsten Seite dargestellt.

Bei dieser Varianzanalyse muss man den Faktor A praktisch nicht interpretieren. Hinter ihm versteckt sich der Faktor „Qualität der Aufsätze", nur dass diesmal die Aufsätze 2 + 3 als mittellange Aufsätze zusammengefasst wurden. Dass die Beurteilung der Aufsätze sich auch im Hinblick auf die Länge unterschied (*F=490,6; df=2/350; p<.001*), wird durch die Überlegung einsichtig, dass Aufsatzlänge und Qualität hoch korreliert waren. Der längste Aufsatz war zugleich der beste und der kürzeste der schlechteste. So ist in diesem Falle nicht unmittelbar entscheidbar, ob wirklich die Länge oder die Qualität des Aufsatzes die Beurteilung beeinflusst hat. Beides wäre denkbar.

Quelle	SAQ	df	Varianz	F	p <	Effekt%
Zw. A (Länge)	3812768	2	1906384	490,56	.001	72,29
Zw. B (Fehler)	104040	1	104040	26,77	.001	1,97
Zw. A*B	2404	2	1202	0,31	n.s.	0,05
Innerhalb	1360156	350	3886			25,76
Total	5279368	355				100,00

Tab. 3: Varianzanalyseergebnisse

Um etwas genauer den Einfluss der Aufsatzlänge abschätzen zu können, wurde eine Kovarianzanalyse angeschlossen, bei der die Qualität der Aufsatzleistung durch Eingabe der Durchschnittsbeurteilungen als Kovariate auspartialisiert wurde. Dabei reduzierte sich der *F-Wert* des Haupteffekts A „Aufsatzlänge" auf *0,318*. Er verschwindet also ganz. Demnach handelt es sich bei dem Haupteffekt des Faktors A „Aufsatzlänge" wohl doch eher um einen Effekt der „Aufsatzqualität"!

Noch interessanter ist die Tatsache, dass die Fehlerhaftigkeit der Texte (Faktor B), damit also die Rechtschreibleistung, das Lehrerurteil signifikant beeinflusste (*F=26.8; df=1/350; p<.001*). Mit ca. 2% Varianzklärung fällt dieser Faktor aber zunächst nicht besonders ins Gewicht. In der Kovarianzanalyse steigt dann der F-Wert des Haupteffekts B „Fehlerhaftigkeit" an auf *27.3*, was dort einer Varianzklärung von 6,98% entspricht. Insofern kann man den Effekt, den die Variation der Fehler auf die Aufsatzbeurteilung hat, nicht vernachlässigen. Fast 7% signalisieren eben doch einen beachtenswerten Einflussfaktor!

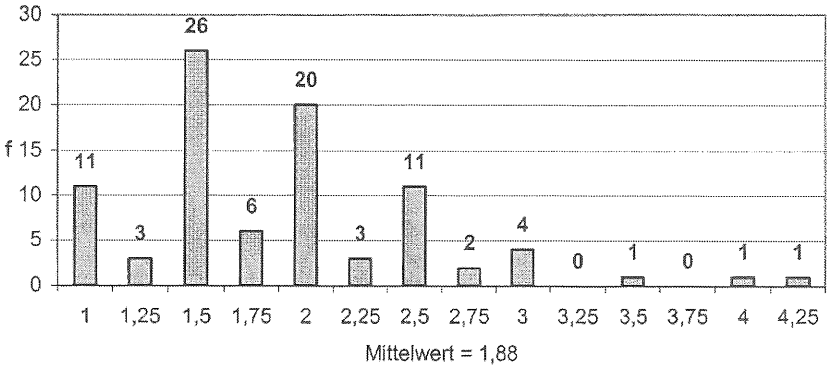
Besonders deutliche Unterschiede in den Beurteilungen ergaben sich vor allem, wie weiter oben bereits dargestellt, wenn der beste Aufsatz mit wenig Fehlern und der

schlechteste zusätzlich mit vielen Fehlern zu beurteilen war. Hier kommt es praktisch zu einer Art „Kontrasteffekt", ähnlich dem, der von Birkel (1978) im Zusammenhang mit mündlichen Prüfungen beschrieben wurde. Die Beurteilungsdifferenz, die auf die Fehlerhaftigkeit der Texte zurückgeführt werden kann, liegt bei 0.35 Notenstufen. Man kann also sagen, dass die Aufsätze mit vielen Fehlern im Schnitt um etwa ⅓ Notenstufe schlechter beurteilt wurden (2.89 bei wenig Fehlern; 3.24 bei vielen Fehlern).

Urteilsübereinstimmung

Hier geht es jetzt um die Frage, ob die beurteilenden Lehrer bei den vier Aufsätzen einigermaßen ähnliche abschließende Beurteilungen abgegeben haben, oder ob die Beurteilungen weit auseinander klaffen. Dazu soll die Häufigkeit der Noten dargestellt werden, die den einzelnen Aufsätzen erteilt wurden.

Abb. 3: Aufsatz 1, alle Beurteilungen



Bei Aufsatz 1 stellt man fest, dass die erteilten Noten von der glatten 1 bis zur 4- (4,25) reichen. Der Mittelwert liegt bei 1,88, die Standardabweichung bei 0,66. Alle Beurteilungen im Bereich $M \pm S$, dem Konfidenzintervall[6], sind gut nachvollziehbar, bei diesem Aufsatz also die Zensuren zwischen 1,22 (1-) und 2,54 (2-3). Immerhin 11 Lehrer erteilten eine glatte 1 und lagen damit außerhalb dieses Konfidenzintervalls. Ebenfalls vergaben zwei Lehrer die 3+ und vier noch die Note 3. Große Rechtfertigungsprobleme dürften aber die Lehrer bekommen, die eine 3-4, eine 4 und sogar eine 4- gaben. Leider waren in diesen Fällen die Noten nicht ausführlich begründet.

Waren bei Aufsatz 1 generell schon große Beurteilungsunterschiede erkennbar, so wäre es nun interessant zu prüfen, ob sich die Unterschiede möglicherweise reduzieren, wenn man die Beurteilungen betrachtet, die bei Vorliegen von vielen oder wenigen Fehlern abgegeben wurden. Aus Abb. 4 geht hervor, dass die Beurteilungsun-

---

6 Als Konfidenzintervall bezeichnet man den Bereich Mittelwert ± Standardabweichung. Bei einer Normalverteilung liegen in diesem Bereich gut ⅔ aller Beurteilungen.

terschiede sich kaum reduzieren. Die beste und schlechteste Note wurden von Lehrkräften erteilt, die den Aufsatz mit wenig Fehlern zu beurteilen hatten. Andererseits erteilten auch Lehrkräfte bei Vorliegen vieler Fehler durchaus auch die Note 1. Trotzdem ist sichtbar, dass in letzterem Falle deutlich häufiger nur die mittelmäßigen Noten 2, 2,5 und 3 erteilt wurden[7]. Der Zusammenhang zwischen Fehlerhaftigkeit und Note liegt bei diesem Aufsatz bei $r = -.309$ und ist auf dem 1%-Niveau signifikant. Die Rechtschreibung determiniert hier 9.55% der Notenvarianz[8].

Abb. 4: Aufsatz 1

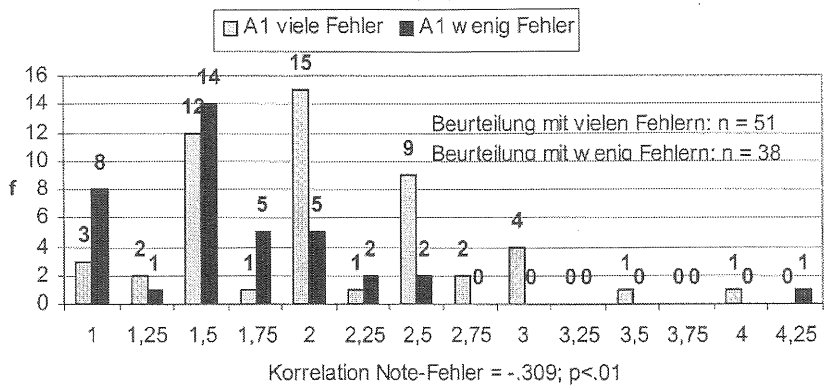
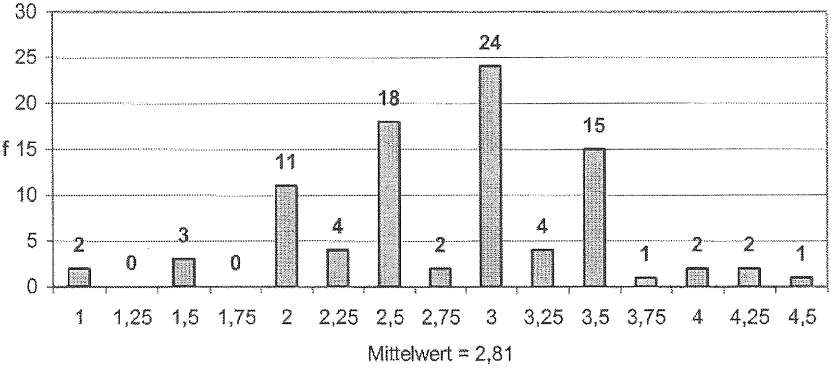


Abb. 5: Aufsatz 2, alle Beurteilungen

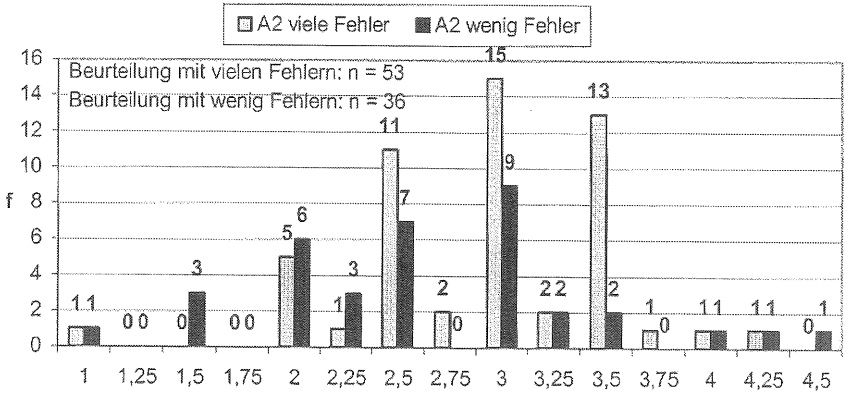


7 Bei der Interpretation der Graphiken, die die Häufigkeiten der Noten bei Vorliegen von vielen oder wenig Fehlern darstellen, ist zu bedenken, dass sich bei jedem Aufsatz die Anzahl der Lehrer sich deutlich unterscheiden kann. In diesem Falle haben 51 Lehrer den Aufsatz mit vielen Fehlern und 38 mit wenig Fehlern zu beurteilen gehabt.

8 Der Determinationskoeffizient errechnet sich als $D = r^2(*100\%)$.

Beim Aufsatz 2 (Abb. 5) ist sofort sichtbar, dass die meisten Lehrer die Note 3 erteilten. Sie liegt der durchschnittlich beurteilten Qualität des Aufsatzes ($M = 2,81$; $S = 0,69$) recht nah. Alle Beurteilungen innerhalb des Konfidenzintervalls $M \pm S$, also zwischen den Noten 2,12 und 3,50, sind sicher jederzeit vertretbar, besonders wenn man bedenkt, dass die mehr oder minder gute Rechtschreibfähigkeit der Schüler in die Endnote für den Aufsatz einging. Immerhin 16 Lehrer griffen aber zur Note 2 oder besser und lagen mit ihrer Beurteilung ebenso außerhalb des Konfidenzintervalls wie die 6 Lehrer, die eine schlechtere Note als die 3,5 erteilt hatten. Die Spannweite der Beurteilungen ist doch sehr groß!

Abb. 6: Aufsatz 2



Korrelation Note-Fehler = -.209; p<.05

Wieder könnte man einwenden, dass der große Streubereich bei den Zensuren nur auf die Tatsache zurückzuführen wäre, dass ja die Lehrer auch wieder den Aufsatz 2 zum Teil mit vielen, zum Teil mit wenigen Fehlern zu beurteilen hatten. Dass der Einwand nicht sticht, zeigt Abb. 6. Gerade bei den Lehrern, die den Aufsatz mit nur wenig Fehlern zu beurteilen hatten, reichte die Notenstreuung von der glatten 1 bis zur 4,5. Auf der anderen Seite beurteilte selbst eine Lehrkraft, die den Aufsatz mit vielen Fehlern zu beurteilen hatte, diesen mit einer glatten 1. Hier reicht die Streubreite bis zur Note 4- hinauf. Der Zusammenhang zwischen der Fehlerhaftigkeit des Aufsatzes und der Gesamtnote lag hier übrigens bei $r = -.209$ und war auf dem 5%-Niveau signifikant. Die Klärung der Notenvarianz beträgt in diesem Falle 4,37%.

Beim Aufsatz 3 (s. Abb. 7) finden wir fast das gleiche Bild wieder, nur dass die Spannweite der Beurteilung diesmal von der 1 bis zur glatten 5 reicht und dabei mit ganzen vier Notenstufen ein Maximum erreicht! Der Aufsatz ist bei einer Länge von 238 Wörtern relativ ausführlich, inhaltlich gesehen vielleicht nicht besonders spannend, aber doch so geschrieben, dass die Reizwörter sinnvoll integriert wurden[9]. Die 3, die die meisten Lehrer diesem Aufsatz erteilten, ist wirklich gut nachvollziehbar,

9 Die Anwendung des Züricher Analyserasters auf diesen Aufsatz würde klar machen, dass doch einige der Kriterien deutlich erfüllt sind.

reicht doch Konfidenzintervall diesmal von der 2,55 bis zur 3,91! Dass darüber hinaus auch noch ein gewisser subjektiver Ermessensspielraum bei der Beurteilung zu berücksichtigen ist, wird auch jeder zugestehen. Zwischen den Noten 2,5 und 3,5 lässt sich jede Beurteilung rechtfertigen. Selbst eine Spannweite der Noten von der 2 bis zur 4 wäre unter Berücksichtigung der eventuell ja sehr unterschiedlich guten Rechtschreibleistung durchaus zu vertreten. Benotungen besser als 2 und schlechter als 4 fallen aber doch aus dem Rahmen, und das sind immerhin 7 Beurteilungen! Es ist besonders schwer nachvollziehbar, warum derselbe Aufsatz mit einer 1 oder aber ebenso gut auch mit einer 5 beurteilt werden könnte! Hier wäre es sehr sinnvoll, wenn man die Argumentationsfiguren der beurteilenden Lehrer erfassen könnte, was bei dieser Untersuchung aber leider nicht möglich war.

Abb. 7: Aufsatz 3, alle Beurteilungnen

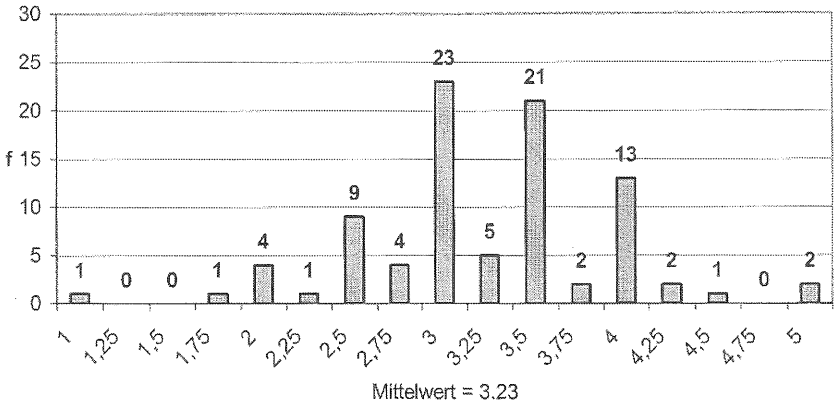
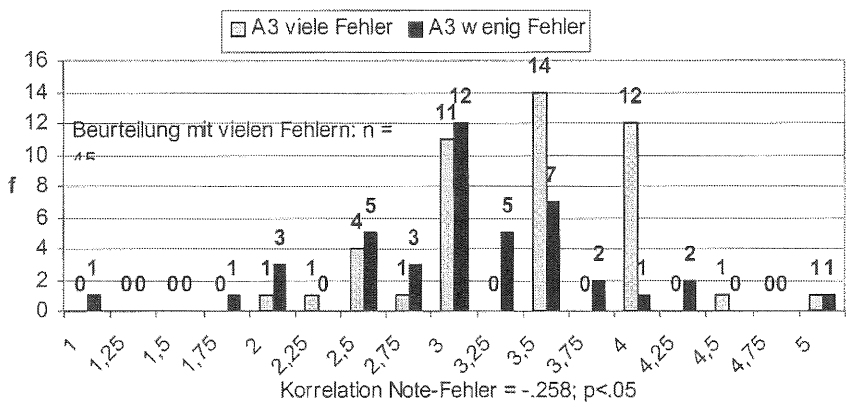


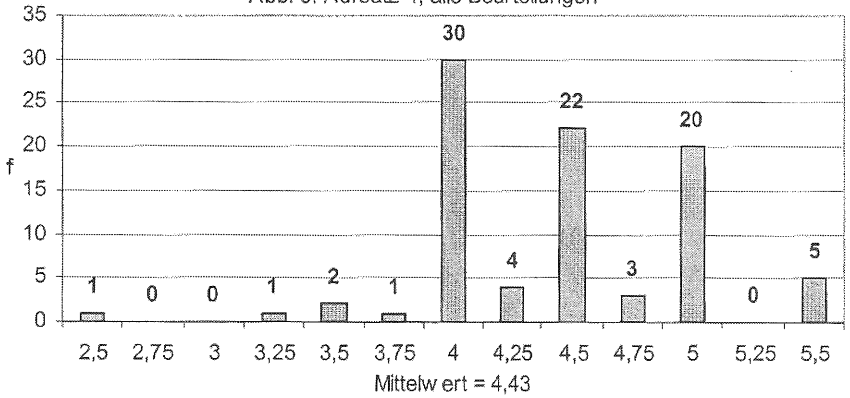
Abb. 8: Aufsatz 3



Erneut fällt auf, dass bei Aufsatz 3 die Spannweite der Beurteilungen nur bedingt reduziert wird durch die Aufgliederung nach Beurteilungen mit vielen oder wenig

Fehlern (s. Abb. 8). Bei Beurteilung mit wenig Fehlern reicht die Spannweite von der 1 bis zur glatten 5, umfasst also vier Notenstufen! Nur bei der Beurteilung mit vielen Fehlern reduziert sich die Spannweite um eine Notenstufe auf den Bereich von der 2 bis zur 5. Wieder ist auch die Versetzung der beiden Verteilungen gegeneinander gut erkennbar. Mit vielen Fehlern werden schon insgesamt auch häufiger etwas schlechtere Noten vergeben. Von daher ist auch die Höhe des Zusammenhanges nicht erstaunlich. Mit $r = -.258$ ist sie auf dem 5%-Niveau signifikant und deutet auf eine Varianzaufklärung von 6,66% hin.

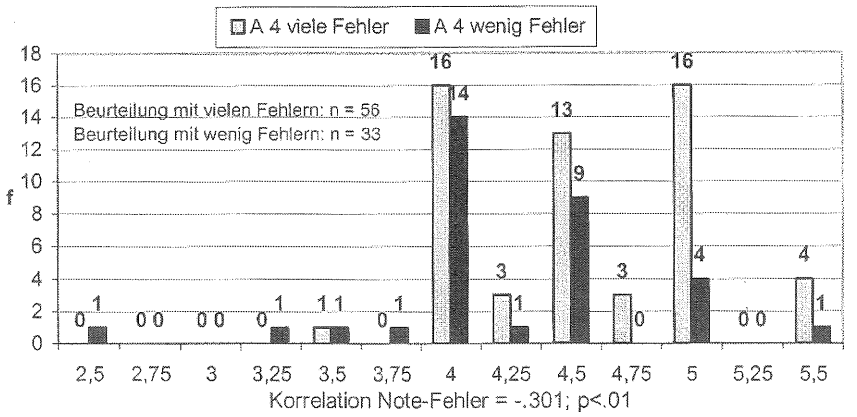
Abb. 9: Aufsatz 4, alle Beurteilungen



Bei Aufsatz 4 fällt zunächst auf, dass die Streubreite der erteilten Noten geringer ausfällt als bei den anderen Aufsätzen. Trotzdem erreicht sie auch hier noch drei Notenstufen. Aufsatz 4 wurde nicht nur von fast allen Lehrern deutlich als der schlechteste identifiziert, wie aus Abb. 9 zu ersehen ist, er war auch im Original mit einer 4- der am schlechtesten beurteilte. Die Lehrer, die bei diesem Aufsatz eine 4-5 erteilten, lagen am nächsten an der Durchschnittsbeurteilung von *4.43*. Es lässt sich wieder ein Konfidenzintervall zwischen den Werten 3,88 und 4,97 eröffnen, in der alle Beurteilungen als vertretbar erscheinen. In diesem Falle liegen je 5 Beurteilungen oberhalb und unterhalb des Konfidenzintervalls, wenn man die 4,97 als glatte 5 berücksichtigt. Bei der Lehrerin oder dem Lehrer, die/der für diesen Aufsatz sogar eine 5-6 erteilte, wäre natürlich die Begründung für diese Note sehr interessant gewesen. Ebenso sehr hätten die Gründe interessiert, die angeführt worden wären, um die Note 2-3 verständlich zu machen. Möglicherweise hätten sich beide auf völlig unterschiedliche Aspekte der Aufsatzleistung bezogen. In dieser Richtung würde man sicher in Zukunft genauer nachfragen müssen.

Abb. 10 zeigt, dass die Spannweite der Noten bei Beurteilung mit wenig Fehlern erneut am größten ist. Sie reicht von der 2,5 bis zur 5,5! Eher erwartungsgemäß ist hier, dass bei vielen Fehlern die Noten erst bei der 3,5 beginnen. Auch der Zusammenhang zwischen Fehlerhaftigkeit und Noten liegt in dem Bereich, der auch schon bei den anderen Aufsätzen beobachtbar war. Mit $r = -.301$ ist er auf dem 1%-Niveau signifikant und klärt 9,06% der Notenvarianz.

Abb. 10: Aufsatz 4



Die Notenverteilungen der vier Aufsätze fielen nicht nur im Hinblick auf die Streubreiten der Noten unterschiedlich aus, sondern auch generell bei den Streuungstendenzen. Diese werden dokumentiert mit der Angabe der Standardabweichungen.

Aufsatz	Standardabweichung
1	0.655
2	0.686
3	0.681
4	**0.540**

F = 1.473: p <
F = 1.616: p <
F = 1.594: p <

Tab. 4: Streuungsmaße für die 4 Aufsätze

Hier ist zu bedenken, dass die Einigkeit in der Beurteilung der Aufsätze umso größer einzuschätzen ist, je kleiner der Wert für die Standardabweichung ausfällt. Im Falle dieser Untersuchung ist zu erkennen, dass besonders deutlich die Streuung bei Aufsatz 4 geringer ausfällt als bei den anderen Aufsätzen. Die Überprüfung der Streuungsunterschiede mit dem F-Test ergab, dass tatsächlich die Streuung bei Aufsatz 4 signifikant geringer war als bei den anderen Aufsätzen, die sich wiederum untereinander nicht signifikant unterschieden. Das bedeutet, dass es wohl leichter ist, bei einer schlechten Leistung ein größeres Maß an Übereinstimmung zu erreichen als bei mittelmäßigen oder guten Leistungen.

Dieses Ergebnis steht im Einklang mit Überlegungen von Birkel (1978), der auch bei der Beurteilung mündlicher Reifeprüfungen schon vermutete: „Die Validität der Beurteilungen verringert(..) sich (..), je schwieriger es für die Beurteiler (ist), die (..)Leistungen in ihr Referenzsystem einzuordnen und sie mit einer Zensur zu bezeichnen."(S. 200) Genau das ist aber bei mittelmäßigen Leistungen der Fall. Hier ist die subjektive Bewertungsunsicherheit am größten. „In solchen Fällen sucht der Beurteiler mehr oder weniger bewusst nach Anhaltspunkten, die ihm eine Einord-

nung in sein Referenzsystem erleichtern können, und greift dabei u.U. auch auf solche zurück, die eigentlichen nichts mit der gezeigten Leistung zu tun haben." (S. 200) An dieser Stelle bieten dann z.B. die Rechtschreibleistung oder die Geschlechtszugehörigkeit Anhaltspunkte, die die Beurteilung erleichtern nach dem Motto: Wer wenig Fehler macht, ist auch der bessere Aufsatzschreiber (Halo-Effekt, s. Schwarzer/Schwarzer 1977, S. 13), oder Mädchen können besser Aufsätze schreiben als Jungen, weil sie sprachlich gewandter sind (implizite Theorie, s. Schwarzer/Schwarzer 1977, S. 27f).

6 Abschließende Bewertung

Bei der Planung dieser kleinen Untersuchung war man einhellig der Meinung, dass die Beurteilungsunterschiede bei den zu beurteilenden Aufsätzen eher gering ausfallen würden. Als alle Teilergebnisse zusammengefasst worden waren, stellte sich allerdings Erstaunen ein. Dass die Beurteilungsunterschiede insgesamt dann doch so groß werden würden, damit hatte niemand gerechnet. Im Grunde war sogar erwartet worden, dass eine deutlichere Übereinstimmung bei den abgegebenen Beurteilungen möglich gewesen wäre, weil durch die unterschiedliche Qualität der Aufsätze, die sich ja auch in den unterschiedlichen Durchschnittsbeurteilungen dokumentiert, bereits eine Art Normierung vorgegeben war. Bei der Vorlage qualitativ ähnlicher Aufsätze aus dem mittleren Notenbereich hätte die Streuung der angegebenen Noten sogar noch deutlich größer ausfallen können. Umso bedeutsamer sind die trotzdem noch gefundenen Notenstreuungen!

Halten wir also fest, dass sich im Prinzip die Ergebnisse von Weiss (1965, 1966) replizieren ließen. Die Tatsache, dass Ingenkamp diese Mängel 1971 bereits sehr deutlich dargestellt hat, und sie auf Befragen hin der Lehrerschaft sehr wohl bekannt sind, ändert nichts daran, dass diese Mängel nach wie vor wirksam sind. Es gibt wohl kaum etwas Schwierigeres für Lehrer, als sich bei einer abschließenden globalen Einschätzung der Qualität eines Aufsatzes auf ein vergleichbares Urteil zu einigen. Mit dem Züricher Textanalyseraster läge ein Instrumentarium vor, das nicht nur empirisch gefunden wurde, sondern auch fachdidaktischen Kriterien genügt.

Seine Übersetzung auf verschiedene Schulstufen und –arten und seine Anpassung darauf sowie weitere empirische Untersuchung zu seiner Validität stehen allerdings noch aus. Das mag damit zu tun haben, dass in der gegenwärtigen Aufsatzdidaktik mehr der Prozess als das Produkt gesehen wird. Dieser Einwand verkennt aber, dass auch bei Beurteilungen, die nicht nur auf ein Endprodukt (den bekannten „Schulaufsatz") eine Note vergeben, sondern auch den je besonderen Lernfortschritt beim Schreiben mit berücksichtigen, jeweils ein Textzustand als zu beurteilendes Produkt vorliegt. Solange ganz offensichtlich dafür ein geschulter Blick fehlt, hilft die Umorientierung der Textdidaktik vom Produkt zum Prozess dem lernenden Schüler/ der lernenden Schülerin wenig.

Literatur

Baurmann, J. (1975) Aufsatzbenotung und Reihenfolgeeffekt. Beeinflusst die Reihenfolge im Beurteilungsvorgang die Aufsatzbenotung? In: *Psychologie in Erziehung und Unterricht*, 22, S. 181-185

Baurmann, J. (1977)[7] Der Einfluss von Auswertungsbedingungen, Vorinformationen und Persönlichkeitsmerkmalen auf die Benotung von Deutschaufsätzen. In: Ingenkamp, K. (Hrsg.): *Die Fragwürdigkeit der Zensurengebung*. Beltz. Weinheim, S. 117-130

Baurmann, J. & Gier, E.-M. (1985) Zur Aufsatzbeurteilung: eine Auswahlbibliographie. In: *Praxis Deutsch*, 12 (72) S. 10-17

Beck, O. (1974) *Kriterien zur Aufsatzbeurteilung*. Von Hase und Koehler: Mainz

Beck, O. & Hofen, N. (1991) *Aufsatzunterricht Grundschule*. Schneider: Hochgehren

Beck, O. & Payrhuber, F,J. (Hrsg.) (1975) *Aufsatzbeurteilung heute. Problematik, Diagnose, Therapievorschläge*. Herder: Freiburg

Birkel, P. (1978*) Mündliche Prüfungen. Zur Objektivität und Validität der Leistungsbeurteilung*. Kamp Verlag: Bochum

Diederich, P.B.; French, J.W. & Carlton, S.T.: (1961) *Factors in Judgement of Writing Ability*. (Research Bulletin RB 61-15) Princeton, NJ: Educational Testing Service

Finlayson, D. S. (1977)[7] Die Zuverlässigkeit bei der Zensierung von Aufsätzen. In Ingenkamp, K. (Hrsg.): *Die Fragwürdigkeit der Zensurengebung*. Beltz: Weinheim, S. 90-103

Fliegner, J. (1974) *Normative Beurteilung von Schüleraufsätzen mit Hilfe eines Punktsystems*. Düsseldorf

Godshalk, F.I.; Swineford, F. & Coffman, W.E. (1966) *The Measurement of Writing Ability*. New York: College Entrance Examination Board

Grzesik, J. & Fischer, M. (1984) *Was leisten Kriterien für die Aufsatzbeurteilung? Theoretische und praktische Aspekte des Gebrauchs von Kriterien und der Mehrfachbeurteilung nach globalem Eindruck*. (Forschungsberichte des Landes Nordrhein-Westfalen Nr. 3192) Opladen: Westdeutscher Verlag

Ingenkamp, K. (Hrsg.) (1971) *Die Fragwürdigkeit der Zensurengebung*. Weinheim: Beltz, 7. überarbeitete und ergänzte Auflage 1977

Ingenkamp, K. (1989) Die diagnostische Problematik des Aufsatzes als Prüfungsinstrument und die Bemühungen zur Verbesserung der Auswertungsqualität. In: Ingenkamp, K.: *Diagnostik in der Schule*. Weinheim: Beltz

Lehmann, R.H. (1987) Reliability and Generalizability of Ratings of Compositions. In: Degenhart, R.E. (Ed.): *Assessment of Student Writing in an International Context*. Jyväskylä: Institute for Educational Research. Publication Series B: Theory into Practice 9, pp. 141-152

Lehmann, R.H. (1988) Zuverlässigkeit und Generalisierbarkeit von Aufsatzbewertungen. In: *Empirische Pädagogik* (2) S. 349-365

Lehmann, R.H. (1990a) Aufsatzbeurteilung – Forschungsstand und empirische Daten. In: Ingenkamp, K. & Jäger, R.S. (Hrsg.): *Tests und Trends*. Jahrbuch der Pädagogischen Diagnostik, Bd. 8, Weinheim: Beltz, S. 64-94

Lehmann, R.H. (1990b) Reliability and Generalizability of Ratings of Compositions. In: *Studies in Educational Evaluation* (16) pp. 501-512

Lehmann, R.H. (1993) Rating the Quality of Student Writing – Findings from the IEA Study of Achievement in Written Composition. In: Huhta, A.; Sajavaara, K. & Takala, S. (Eds.): *Language Testing: New Openings*. Jyväskylä: Institute of Educational Research, pp. 186-204

Lehmann, R.H. (1994a) Research on National and International Writing Assessments: Contributions from the Hamburg Study of Achievement in Written Composition. In: Ansorge, R. (Hrsg.): *Schlaglichter der Forschung*. Zum 75. Jahrestag der Universität Hamburg 1994. Berlin: Dietrich Reimer Verlag, S. 173-184

Lehmann, R.H. (1994b) Essays, Scoring of. In: Postlethwaite, T.N. & Husén, T. (Eds.): *International Encyclopaedia of Education*. Vol. 4, 2nd ed., Oxford, pp. 2018-2025

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

Osnes, J. (1977)[7] Der Einfluss äußerer Faktoren bei der Aufsatzbeurteilung. In: Ingenkamp, K. (Hrsg.): *Die Fragwürdigkeit der Zensurengebung*. Beltz: Weinheim, S. 131-147

Schwarzer, Ch. & Schwarzer, R. (1977) *Praxis der Schülerbeurteilung*. Kösel: München 1977

Weber, A. (1973) *Dialektik der Aufsatzbeurteilung*. Donauwörth: Auer

Weiss, R. (1965a) *Zensur und Zeugnis*. Haslinger: Wien

Weiss, R. (1965b) Über die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen. In: *Schule und Psychologie*, Heft 9, S. 257-269

Weiss, R. (1966a) Über die Auswirkung bestimmter Einstellungen auf Zensuren. In: *Unser Weg*, S. 166-177

Weiss, R. (1966b) Über die Zuverlässigkeit der Ziffernbenotung bei Rechenarbeiten. In: *Schule und Psychologie*, Heft 5, S. 144-151

Weiss, R. (1977)[7] Die Zuverlässigkeit der Ziffernbenotung bei Aufsätzen und Rechenarbeiten. In: Ingenkamp, K. (Hrsg.): *Die Fragwürdigkeit der Zensurengebung*. Beltz: Weinheim, S. 104-116

Zeiher, H.; Zeiher, H. & Krüger, H. (1979) *Textschreiben als produktives und kommunikatives Handeln*. Bd. 1-3, Stuttgart: Klett

Anschrift des Verfassers: *Dr. Peter Birkel, Fach Pädagogische Psychologie, Pädagogische Hochschule Weingarten, Kirchplatz 2, 88250 Weingarten, E-Mail: birkel@ph-weingarten.de*

Anhang

Aufsätze mit den Reizwörtern: Langeweile - Dachboden – Kleidertruhe

Aufsatz 1: Der geheimnisvolle Dachboden

Ich wollte immer von meiner Mutter etwas von der Zeit, der Kleidung, eben alles erfahren, wie es früher einmal wahr. Es war mir ja auch so langweilig! Nach diesem langweiligen Thema faszinierte mich der Dachboden. Ihn fand ich am tollsten von allen Sachen. Er ist so geheimnissvoll und voller uralter sachen, die ich doch so liebte! Mama wurde einmal furchtbar böse als sie mich an der Tür zum Dachboden sah. (Einen Papa habe ich nicht.) Seitdem suche ich nach einer Gelegenheit hineinzukommen.

Eines Tages ging Mutti zum Einkaufen und ließ mich ganz allein zu Hause. Das war die Geleggenheit! Obwohl es gar nicht nötig war, ging ich auf Zehenspitzen die Treppe hinauf, um Die Ecke und zur Tür hin. Ich machte die Tür auf. Ich sah einen spinnenwebenbehangenen Raum vor mir. Da waren so viel Spinnenwippen, dass ich gar nichts sehen konnte. Arbeitstüchtig, wie immer, holte ich einen Besen und fing mit der Arbeit an. Als erstes machte ich natürlich die Spinnenwippen weg, bei meiner Kehrexpedition entdeckte ich ein altes Klavier, ganz schwarz, alte Bilder, bestimmt von einem Künstler, ein altes Aquarium, wo bestimmt schon viele Lecke hatte, altes Geschirr, einen Korb voller Kreuze aller Art, alte Stofftiere, einen Globus und überall an der Wand standen Regale mit vielen, vielen Büchern! Als ich schon dachte ich hatte alles gesehen, entdeckte ich noch eine ur, ur, uralte True, als ich sie öffnete, knarrte sie laut - da drin waren wunderschöne, alte, ganz alte Kleider, Hosen Röcke, aber vorallem lange Kleider! Und alles was ich entdeckt hatte, auch der Boden war mindestens mit 1 ½ cm Staub bedeckt gewesen.

In dem Moment knallte die Türe, die Treppe rumpelte und in der Tür stand Mutti!

Aber statt sie böse war und mich verdrosch, wie das letzte Mal, lachte sie aus vollem Halse und meinte kichernd: „Komm vorsichtig herunter und sie dich mal im Spiegel an, du Dreckspatz!“ Und wahrhaftig als ich mich im Spiegel sah, musste ich auch lachen. Alle meine Kleider wahren grau mein Gesicht und auch meine Haare auch. Ich sah aus wie eine kleine, total graue Oma, Ein Marsmensch! Danach wusch ich mich noch und Mutti sagte: „Wenn du willst können wir zusammen dort aufräumen, dann hast du ein Zimmer ganz allein für dich zum Spielen und dir wird nicht mehr langweilig sein.

Note: 2+

Aufsatz 2: Was ist los?

Heute war ein kalter Tag und meine Freundin ist nicht zuhause. So musste ich eben alleine spielen, aber: das ist so langweilig.

Doch da hatte ich eine Idee. Ich stieg die Treppe hinauf zum Dachboden. Nun saß ich oben und sah die Kleidertruhe. So dachte ich: „Es ist zwar nicht Fasching, aber ich könnte mich doch verkleiden.“ Ich schlich zur Truhe und hob den Deckel hoch, und schaute mir die Kleidungsstücke genau an. Es waren alte Röcke, bunte Hemden und viele Hosen. Doch plötzlich hörte ich meinen Namen schreien. Ich sprang auf, doch in dem Moment riss ich die Truhe um. Ich rannte die Treppe hinunter und lief zu meiner Mutter und fragte: „Was ist los?“ Mutti antwortete: „Ich wollte nur wissen ob du da bist.“ So ging ich

wieder die Treppe hinauf und sah mir das Durcheinander an und sagte: „Jetzt kann ich aufräumen, da wird es mir nicht mehr langweilig."

Seit dem was passiert ist ging ich nie wieder auf den Dachboden zur Kleidertruhe.

Note: 3+

Aufsatz 3: Ein Wirbel um die Klamotentruhe

Eines Tages hate Philipp Langeweile und wuste nicht was tun. Aber dan viel im ein das er eine große Truhe gesen hat er ging einfach Leise nach oben.

Dan suchte er die Truhe auf dem Dachboden er sah sie nach einer weile machte er die Truhe ganz langsam auf und sah das da Kleidertruhe ja ist. Er schaute hinein und schaute ob was für ihn darin ist er suchte ganz lange bis er etwas gefunden hate. Da fand er einen schöhnen Pullofer aber er zeihte ihn ihn nicht seiner Mutter weil er will auch mal ein geheimnis haben. Er blieb lange oben er fand noch eine schöhen sgeiterschorts und einen Tisch und noch vieles anderes sachen. Abens ging er wieder runter und fragte sich ob er seiner Mutter doch sagen sol aber er sagte in seinen gedanken: „Die Kleidertruhe muss ich fersteken aber wie mach ich das nur da mus ich mir etwas einfallen lasen." Er dachte nach was er mit der truhe machen sollte da fiel im etwas ein er hat in seinem Zimer eine kamer wo er sie ferschteken kann. Da ging er wieder zu der Dachboden zum die Kleidertrtuhe holen. Jetzt ist es ales sein in der Truhe. Dan wo er die Truhe in sein Zimer bepracht hat da war in nicht mehr Langeweilig.

Abends im Bett dachter imernoch an die Kleidertruhe. Wo er eingeschlafen war träumte er dafon.

Note: 3-

Aufsatz 4: Es war ein schöner Morgen

Es war ein schöner Morgen etwas zu spielen doch leider hate ich Schule nach der Schule ging ich mit meinen Freundinnen zum spielen. Anna hate uns eingeladen.

Am Anfanang haten wir Langeweile aber dan viel Tanja etwas ein. Fersteke zu spielen als erstes Zelte ich hate aber beide gleich gefunden, Anna dan Tanja. Danach Zelte Anna und Tanja lief zum Dachboden hinauf da rief sie uns und wir gingen auch hinauf, ganz schön gruselig. Und da fanden wir eine Kleidertruhe.

Aber da rief uns Annas Mutter und wir mußten gehen.

Note: 4-