

**Vergleichsarbeiten Klasse 10/Frühjahr 2004.**  
**Bewertungskonstanz bei der Schreibaufgabe im Fach Deutsch.**  
**Erst- und Zweitkorrektur im Vergleich**

**Überblick**

<b>0</b>	<b>Zentrale Ergebnisse im Überblick</b>	<b>1</b>
<b>1</b>	<b>Fragestellung, Anlage und Durchführung der Untersuchung</b>	<b>2</b>
<b>2</b>	<b>Ergebnisse: Bewertungskonstanz berlinweit und differenziert nach Schularten</b>	<b>8</b>
2.1	Die zehn Bewertungskategorien der Schreibaufgabe	9
2.1.1	Die Ergebnisse der Zweitkorrektur berlinweit	
2.1.2	Ergebnisse differenziert nach Schularten	
2.2	Die Formalfehler	22
2.2.1	Die Ergebnisse der Zweitkorrektur berlinweit	
2.2.2	Ergebnisse differenziert nach Schularten	
<b>3</b>	<b>Resümee</b>	<b>29</b>
3.1	Zusammenfassung	29
3.2	Schlussfolgerungen	35
<b>ANHANG</b>		<b>38</b>



## 0 Zentrale Ergebnisse im Überblick

Von 746 Deutschvergleichsarbeiten/Klasse 10 aus dem Frühjahr 2004 wurde die Schreibaufgabe einer Zweitkorrektur unterzogen und anschließend die Ergebnisse von Erst- und Zweitkorrektur verglichen, um die Bewertungskonstanz zu untersuchen. Nachstehend wichtige Ergebnisse:

1. *Wie weit bleibt beim Übergang von der Erst- zur Zweitkorrektur die Rangordnung der Arbeiten erhalten? Bleiben die beim ersten Mal besseren oder schlechteren Arbeiten auch beim zweiten Mal die besseren oder schlechteren?*

Auf diese Frage gibt es nur ein eingeschränktes Ja, wobei es Unterschiede gibt je nachdem, aus welcher Schulart die Arbeiten stammen: Jene aus der Hauptschule weisen die höchsten (korrelativen) Zusammenhänge zwischen Erst- und Zweitkorrektur auf, solche aus dem Gymnasium die niedrigsten. Die Ähnlichkeit nimmt bei den Gymnasialarbeiten auch dann nicht zu, wenn der Zweitkorrektor ebenfalls aus dem Gymnasium kommt - sie ist sogar kleiner als bei schulartdifferenter Korrektur.

2. *Wie groß sind insgesamt die Unterschiede zwischen Erst- und Zweitkorrektur unabhängig davon, ob die Abweichung nach unten oder nach oben erfolgt?*

Werden Abweichungen von  $\pm 20\%$  toleriert, d.h. von  $\pm 3$  Punkten bei insgesamt 15 Punkten, dann liegen rund zwei Drittel aller Arbeiten in diesem Toleranzbereich. Dies gilt auch dann, wenn die beiden Bewertungsbereiche *Inhalt* ( $\pm 2$  von 10 Punkten) und *Darstellung* ( $\pm 1$  Punkt von 5 Punkten), die als Korrekturkriterien vorgegeben waren, getrennt betrachtet werden. Gleichwohl gibt es eine (teils schulartspezifische) Tendenz, die Bewertungskategorien der *Darstellung* heterogener anzuwenden.

Die größten Abweichungen finden sich bei den Arbeiten aus dem Gymnasium und der Realschule, die geringsten bei jenen aus der Hauptschule. Schulartidentische Korrektur führt darüber hinaus beim Gymnasium zu größeren Abweichungen als schulartdifferente.

3. *Führt die Zweitkorrektur tendenziell zu schlechteren Ergebnissen als die Erstkorrektur?*

Als Tendenz Ja, als Regel Nein. D.h.: Im Mittel korrigieren die Lehrkräfte etwas milder, die in den betroffenen Klassen auch unterrichten; es gibt aber zahlreiche Fälle, für die das Gegenteil gilt.

### Folgerungen:

- Die landesweit ermittelten Bezugswerte sind verlässlich.
- Bewertungskategorien und Auswertungshinweise sind zu verbessern.
- Verständigung über das Korrigieren ist vor allem im Gymnasium, aber auch in der Realschule zu sichern.
- Das Korrigieren der Arbeiten durch die unterrichtende Lehrkraft führt nicht zu dramatischen Verzerrungen.
- Auf den Bewertungsbereich der Formalfehler (Fehler der sprachlichen Richtigkeit) kann ohne inhaltliche Einbußen verzichtet werden.
- Stichprobenartig ist auch in den nächsten Jahren durch Zweitkorrekturen zu überprüfen, wie weit die Bemühungen um Auswertungsobjektivität erfolgreich waren.

## 1 FRAGESTELLUNG, ANLAGE UND DURCHFÜHRUNG DER UNTERSUCHUNG

### Fragestellung und Relevanz der Untersuchung

Vergleichsarbeiten liefern die Momentaufnahme des Leistungsstandes einer jeden zehnten Klasse in Berlin und bieten somit - sofern ihnen ein differenziertes Aufgabenspektrum eigen ist - eine wichtige diagnostische Perspektive für die Arbeit einer jeden einzelnen Lehrkraft. Der besondere Stellenwert von Vergleichsarbeiten beruht darauf, (landesweite) Bezugswerte zu ermitteln, die es den Schulen und Klassen ermöglichen, sich in der Berliner Schullandschaft zu positionieren.

Ihre zentrale Funktion können Vergleichsarbeiten nur dann erfüllen, wenn ihre Ergebnisse tatsächlich vergleichbar sind, wenn also insbesondere die Auswertungsobjektivität gesichert ist. Eine Maßnahme in diesem Sinne ist die Standardisierung der Arbeiten, die weitgehend aus multiple-choice-Aufgaben bestehen. Eine weitere Maßnahme sind die Auswertungshinweise, die präzise und umfassend sein müssen. Besonders schwer ist Auswertungsobjektivität bei Schreibaufgaben zu erreichen. Im Fach Deutsch ist der zweite Teil der Vergleichsarbeit 2004 eine derartige Schreibaufgabe gewesen und trug mit 15 von insgesamt 50 Punkten erheblich zur Gesamtbewertung bei. Wie stabil die bei der Erstkorrektur vergebenen Punkte sind, sollte durch eine Zweitkorrektur überprüft werden.

Das ist aus zwei Gründen wichtig: Zum Einen werden für die Entwicklung der nächsten Vergleichsarbeiten Hinweise benötigt, wie viel ggf. noch an den Aufgabenformaten zu ändern und vor allem an den Auswertungskriterien und -hinweisen zu präzisieren ist.

Zum Anderen werden bislang die Vergleichsarbeiten von den Lehrkräften korrigiert, die in den betroffenen Klassen unterrichten, wogegen sich einwenden lässt, dies schränke die erforderliche Objektivität in erheblichem Maße ein. Zu fordern sei deshalb, die Arbeiten so schreiben zu lassen, dass aus dem Äußeren die Herkunftsschule nicht erkennbar werde, und sie dann an anderen Schulen bewerten zu lassen. Dies bedeutete - wie sich unschwer erkennen lässt - einen erheblichen Mehraufwand gegenüber der bisherigen Praxis, der nur dann erforderlich werden würde, wenn die Auswertungsobjektivität nicht gewährleistet wäre, also Erst- und Zweitkorrektur bedeutend auseinanderklafften.

Der Vergleich von Erst- und Zweitkorrektur, von dem hier berichtet wird, verfolgt zwei Ziele: Er soll **erstens** eine generelle Einschätzung der Bewertungskonstanz erlauben und **zweitens** die Frage zu beantworten suchen, wie weit schulartspezifische Unterschiede auftreten, und dies in einem zweifachen Sinne:

- Sind die Bewertungen ähnlicher, wenn die Zweitkorrektur aus derselben Schulart als aus einer anderen stammt?
- Gibt es schulartspezifisch größere oder kleinere Differenzen zwischen der Erst- und der Zweitkorrektur, auch wenn beide Korrektoren<sup>1</sup> derselben Schulart angehören? Gibt es also Bewertungsmuster, die über die vorgegebenen Bewertungskriterien hinaus schulartspezifisch wirksam sind?

---

<sup>1</sup> Im gesamten Text umfasst der Begriff Korrektor Korrektor und Korrektorin.

## Vorlauf

Im Frühjahr 2004 fand in Berlin der dritte Durchgang mit schulartübergreifenden Vergleichsarbeiten für die Klasse 10 statt. Dessen Ergebnisse, deren Darstellung einem gesonderten Bericht vorbehalten ist,<sup>2</sup> beruhen auf einer Zufallsstichprobe, die in ihrem Kern aus 80 Klassen der öffentlichen Oberschulen bestand. Tabelle 1.1 zeigt die Schülerzahlen der Stichprobe Deutsch im Vergleich mit den Gesamtzahlen für die Berliner Oberschulen. Es besteht eine hohe Übereinstimmung bei der Aufteilung auf die Schularten in der Grundgesamtheit und in der Stichprobe.

### **1.1 Tabelle: Vergleich der Verteilung von Klassen und Schüler/innen in der Grundgesamtheit und in der Stichprobe Deutsch.<sup>3</sup>**

Schulart	GRUNDGESAMTHEIT				Stichprobe DEUTSCH			
	Klassen		Schüler/innen		Klassen		Schüler/innen	
O	359	24%	9 710	26%	19	25%	492	27%
OH	185	13%	3 515	10%	8	10%	146	8%
OR	263	18%	7 227	20%	15	19%	393	21%
OG	426	29%	11 482	31%	23	30%	597	32%
OBF	237	16%	4 798	13%	12	16%	222	12%
	<b>1 470</b>	<b>100%</b>	<b>36 732</b>	<b>100%</b>	<b>77</b>	<b>100%</b>	<b>1 850</b>	<b>100%</b>

Am ersten Tag der Vergleichsarbeiten ging per Fax ein Brief an die 80 Schulen, in dem sie über die Zugehörigkeit einer ihrer zehnten Klassen<sup>4</sup> zur Stichprobe informiert wurden, deren Ergebnisse in den einzelnen Fächern also zurückzumelden seien. Für die betroffenen Klassen sollte eine Liste angelegt werden, auf der jedem/r Schüler/in eine Nummer zugeteilt wurde. Über die Nummern, die für die Rückmeldebögen in jedem der Fächer zu verwenden waren, sollte sichergestellt werden, dass einerseits die zentrale Auswertung anonym erfolgte und dass andererseits die Ergebnisse pro Schüler/in in jedem der Fächer verglichen werden konnten.

Im Fach Deutsch erging darüber hinaus die Bitte, nicht nur die Ergebnisse zurückzumelden, sondern zugleich einen halben Klassensatz an Deutscharbeiten mitzuschicken, und zwar die Arbeiten mit ungeraden Schülernummern<sup>5</sup>. Damit sollte eine Streuung der Arbeiten über das Leistungsspektrum und eine repräsentative Auswahl erreicht werden. Eine nachdrückliche und nicht immer einfach zu vermittelnde Bitte richtete sich an die korrigierende Lehrkraft: Sie war gehalten, keinerlei Eintragungen auf der Arbeit vorzunehmen, also z.B. auf einer Kopie zu korrigieren oder eine unkorrigierte Kopie zu schicken.

<sup>2</sup> Zu erreichen über die Menüpunkte *Bildung* und *Qualitätssicherung* auf der Internetseite <http://www.senbjs.berlin.de>.

<sup>3</sup> Angaben zur Grundgesamtheit mit Stand vom 5.9.2003. Einbezogen waren nur öffentliche Schulen.

<sup>4</sup> Um welche Klasse es sich handelte, war vorab per Zufall bestimmt worden.

<sup>5</sup> Ungerade, weil dann bei Klassen mit ungeraden Schüleranzahlen eine Arbeit mehr zurückgeschickt wird, als hätte die Vorgabe gelautet, die Arbeiten mit geraden Nummern zurückzuschicken.

## Anlage und Durchführung der Untersuchung

Es wurden 929 Arbeiten eingesandt. Hiervon waren einige nicht weiter verwendbar, da sie Korrekturzeichen und -anmerkungen oder sonstige Hinweise enthielten, so dass schließlich 816 Arbeiten den Ausgangspunkt für das weitere Vorgehen bildeten.

Die Zweitkorrektur sollte ausschließlich an den rund 240 Schulen erfolgen, die nicht in der Stichprobe aus dem Frühjahr waren. Das hatte zwei wichtige Vorteile: Zum Einen konnte aufgrund der großen Anzahl das Korrigieren auf viele Personen verteilt, also die Arbeitsbelastung pro Lehrkraft verringert werden, und zum Anderen war damit weitestgehend sichergestellt, dass die Arbeiten nicht per Zufall an die erstkorrigierende Lehrkraft gerieten.

Im August und September wurden i.d.R. die Arbeiten jeweils einer Klasse, so weit dies möglich war, per Zufall in Sendungen à vier Stück zusammengestellt. Zuweilen ergaben sich Päckchen von drei oder fünf Arbeiten. Arbeiten mehrerer Klassen wurden - bis auf vereinzelte Ausnahmen am Ende der Versandaktion - nicht gemischt, denn es sollte eine Bewertungssituation ähnlich der der Erstkorrektur hergestellt werden, bei der eine Lehrkraft die Arbeiten ein und derselben Klasse miteinander vergleichen kann. Daher blieben knapp 20 Arbeiten unberücksichtigt und wurden nicht verschickt.

Im Begleitbrief an die Schulleiter/innen wurde darum gebeten, die Arbeiten einer oder zweier Deutschlehrkräfte zu geben, die im vorigen Schuljahr an den Vergleichsarbeiten teilgenommen hatten, und dies mit der Bitte, den zweiten, also nur den Schreibeil, anhand derselben Kriterien wie damals zu korrigieren. Zum Eintragen der Ergebnisse war der entsprechende Ausschnitt aus dem seinerzeit verwendeten Erhebungsbogen sowie die relevanten Ausschnitte aus der Vergleichsarbeit Deutsch 2004 beigelegt. Dass Lehrkräfte die Zweitkorrektur durchführen sollten, deren zehnte Klassen an den Vergleichsarbeiten teilgenommen hatten, geschah, um die Einarbeitungszeit in die Materie zu minimieren, denn zur Bewertung konnte auf die Erfahrung aus dem Frühjahr zurückgegriffen werden.

Beim Versand wurde darauf geachtet, die Arbeiten jeweils einer Schulart ungezielt auf alle Schularten zu verteilen; vgl. die Fragestellungen weiter oben. 748 zweitkorrigierte Arbeiten wurden zurückgesandt. Da aus unbekanntem Gründen zu zwei dieser Arbeiten nicht die Bewertungen der Erstkorrektur vorlagen, gehen 746 Arbeiten in die Auswertung ein. Die Spalte *gesamt* der Tabelle 1.2 zeigt, wie sich die 746 Arbeiten auf die einzelnen Schularten verteilen.

**1.2 Tabelle:** **Schulart, aus der die Vergleichsarbeiten Deutsch stammen, und ihre Aufteilung zur Zweitkorrektur auf die Schularten.** (Angegeben sind die absoluten Häufigkeiten und in Klammern darunter die Zeilenprozente; in der Spalte *gesamt* zusätzlich die Spaltenprozente.)

Erstkorrektur	Schulart der Zweitkorrektur					gesamt
	O	OH	OR	OG	OBF	
O	71 (35%)	33 (16%)	32 (16%)	46 (23%)	20 (10%)	202 (100%/ 27%)
OH	7 (13%)	30 (56%)	3 (6%)	8 (15%)	6 (11%)	54 (100%/ 7%)
OR	4 (2%)	33 (19%)	75 (44%)	46 (27%)	14 (8%)	172 (100%/ 23%)
OG	18 (7%)	58 (23%)	60 (24%)	92 (36%)	24 (10%)	252 (100%/ 34%)
OBF	0 (0%)	8 (12%)	4 (6%)	16 (24%)	38 (58%)	66 (100%/ 9%)
gesamt	100 (13%)	162 (22%)	174 (23%)	208 (28%)	102 (14%)	746 (100%/100%)

Den relativen Häufigkeiten aus der rechten Spalte der Tabelle 1.2 stellen wir die entsprechenden Angaben aus der Tabelle 1.1 gegenüber:

Aufteilung bei	der Grundgesamtheit	der Stichprobe Deutsch insgesamt	den Zweitkorrekturen
O	26%	27%	27%
OH	10%	8%	7%
OR	20%	21%	23%
OG	31%	32%	34%
OBF	13%	12%	9%

Eine kleine Überrepräsentanz gibt es bei den Realschulen und den Gymnasien, während im Verhältnis zur Grundgesamtheit und der ursprünglichen Stichprobe Hauptschule und Berufsfachschule etwas unterrepräsentiert sind. Die Abweichungen sind gering; sie schränken die Aussagekraft der ausstehenden Analysen faktisch nicht ein. Ein Problem stellen eher die teilweise geringe Fallzahlen in einigen Zellen der Tabelle 1.2 dar, wenn wir beispielweise einer Frage nachgehen wollen wie: Ist die Abweichung von der Erstkorrektur der Hauptschularbeiten geringer, wenn der Zweitkorrektor aus dem Gymnasium als aus der Realschule kommt? Auch eine eher explorative Antwort auf derartige Fragen sollte erst dann versucht werden, wenn die Zellgröße mindestens 30 beträgt (Faustregel). Dieser Hinweis macht die Grenzen der Auswertung deutlich, die nur bei einer erheblich größeren Anzahl an Arbeiten hinauszu-schieben gewesen wären.

Die Tabelle 1.2 zeigt auch, dass die meisten Arbeiten einer Schulart von Lehrkräften aus eben dieser Schulart korrigiert wurden. So wurden von den 202 Gesamtschularbeiten 35%, nämlich 71, von Lehrkräften aus der Gesamtschule zweitkorrigiert. Dies gilt entsprechend für alle Schularten.

## **Merkmale, die der Analyse zugrundeliegen**

Im Mittelpunkt der Auswertung stehen zwei Blöcke von Merkmalen. Zum einen die zehn Kategorien (Einzelaspekte), anhand derer die Schreibaufgabe bewertet wurde; zum Anderen die Formalfehler (Fehler der sprachlichen Richtigkeit) Orthographie, Grammatik und Interpunktion, die keinen Einfluss auf die eigentliche Bewertung haben sollten, aber für die Benotung zu berücksichtigen waren, sofern die Schule sich entschieden hatte, die Vergleichsarbeit als Klassenarbeit schreiben zu lassen. Unter Bezug auf die Gesamtzahl der Wörter lassen sich Fehlerquotienten berechnen. Angaben zu den Formalfehlern, die zu machen ausdrücklich ins Ermessen der korrigierenden Lehrkraft gestellt wurde, liegen nur von knapp der Hälfte aller Zweitkorrekturen vor.

Die Auswertung wird sich vorrangig den zehn Bewertungskategorien widmen, die auf die Schreibaufgabe (Aufgabe 18 (Teil II) der Deutschvergleichsarbeit) anzuwenden waren. Sie seien stichwortartig aufgelistet:

### **Inhalt**

- 18-1 (I-1): Präzise Wiedergabe der These in eigenen Worten
- 18-2 (I-2): Argumentation für diese Auffassung, z.B. ...<sup>6</sup>
- 18-3 (I-3): Argumentation gegen diese Auffassung, z.B. ...<sup>7</sup>
- 18-4 (I-4): Gewichtung der Argumente  
(z.B., indem Folgen oder Wertmaßstäbe aufgezeigt werden)
- 18-5 (I-5): Formulierung von eigener Position/Meinung als Schlussfolgerung aus der Argumentation

### **Darstellung und Sprachverwendung**

- 18-6 (D-1) : Erkennbare Gliederung/Struktur
- 18-7 (D-2) : Logisch stimmige Gedankenführung (Kohärenz)
- 18-8 (D-3) : Abstrahierende Begrifflichkeit
- 18-9 (D-4) : Variable und sachgerechte Wortwahl
- 18-10(D-5) : Variable und logisch funktionale Satzverknüpfung

Bei den ersten fünf Aspekten konnten maximal zwei Punkte, bei den anderen ein Punkt vergeben werden. Dementsprechend kennt die hier darzustellende Auswertung die Teilpunktzahl *Inhalt* mit maximal zehn Punkten und die Teilpunktzahl *Darstellung* mit maximal fünf Punkten. Insgesamt konnten im zweiten Teil der Vergleichsarbeit maximal fünfzehn Punkte erreicht werden (bei einer Maximalzahl von fünfzig Punkten).

An zweiter Stelle der Auswertung stehen die Formalfehler, die Fehler der sprachlichen Richtigkeit. Im Einzelnen:

---

<sup>6</sup> In den Hinweisen für die Lehrkräfte folgen an dieser Stelle Beispiele im Hinblick auf das konkrete Thema der Schreibaufgabe "Der schönste Beruf der Welt": Motivation, persönliche Eignung, öffentliches Ansehen, Gestaltungsfreiheit, schnell Geld verdienen, abwechslungsreiche Aufgaben und abwechslungsreiche Arbeitsorte.

<sup>7</sup> Vgl. vorhergehende Fußnote. Die Beispiele für dieses Bewertungskriterium lauteten: Numerus Clausus, fehlende Ausbildungs- oder/und Arbeitsplätze, ungünstige Zukunftsperspektive, fehlende Voraussetzungen (Abschlüsse, Fähigkeiten).

- Fehler
- der Orthographie
  - der Grammatik
  - der Interpunktion.

Zusätzlich waren die korrigierenden Lehrkräfte gebeten worden, auch

- die Gesamtzahl der Wörter

anzugeben, so dass sich aus diesen Angaben ein

- Fehlerquotient "Gesamtzahl aller Fehler/Gesamtzahl der Wörter"

berechnen ließ.

**Hinweis.** Verwendet werden die bislang in Berlin üblichen Abkürzungen für die Schularten:

O: Gesamtschule

OR: Realschule

OH: Hauptschule

OG: Gymnasium

OBF: Berufsfachschule.

## 2 ERGEBNISSE: BEWERTUNGSKONSTANZ BERLINWEIT UND DIFFERENZIERT NACH SCHULARTEN

Nachstehender Abschnitt 2.1 thematisiert die zehn Bewertungskategorien, Abschnitt 2.2 die Formalfehler. Jeder der beiden Abschnitte ist wiederum zweigeteilt: Zunächst werden die berlinweiten Ergebnisse dargestellt, dann schulartspezifische Differenzierungen vorgenommen.

Innerhalb dieser Teilabschnitte folgt die Analyse jeweils einem Dreischritt, der sich anhand dreier Fragen beschreiben lässt:

1. *Wie weit bleibt beim Übergang von der Erst- zur Zweitkorrektur die Rangordnung der Arbeiten erhalten? Bleiben die beim ersten Mal besseren oder schlechteren Arbeiten auch beim zweiten Mal die besseren oder schlechteren?*

Der statistische Zugriff erfolgt über die Bestimmung von Korrelationskoeffizienten, die das Ausmaß derartiger Zusammenhänge quantifizieren.<sup>8</sup>

2. *Wie groß sind insgesamt die Unterschiede zwischen Erst- und Zweitkorrektur unabhängig davon, ob die Abweichung nach unten oder nach oben erfolgt?*

Die Resultate von Erst- und Zweitkorrektur sollten im Idealfall übereinstimmen. Jede Abweichung verletzt die angestrebte Bewertungskonstanz unabhängig davon, ob bei der Zweitkorrektur mehr oder weniger Punkte als bei der Erstkorrektur vergeben werden. Die hier interessierende statistische Größe ist demnach der Absolutbetrag, um den sich Erst- und Zweitkorrektur unterscheiden, also die vorzeichenlose Differenz.

3. *Führt die Zweitkorrektur tendenziell zu besseren oder zu schlechteren Ergebnissen als die Erstkorrektur?*

Hinter dieser Frage steht die zuweilen geäußerte Vermutung, dass Lehrkräfte, die in der Klasse unterrichten, deren Arbeiten sie korrigieren, zur Milde neigen. Antwort auf die Frage liefert die (gerichtete) Differenz "Zweitkorrektur minus Erstkorrektur", die positiv ist, falls die Zweitkorrektur zu einem besseren Ergebnis führt, negativ hingegen, falls beim zweiten Mal weniger Punkte als beim ersten Mal vergeben werden.

---

<sup>8</sup> Korrelationskoeffizienten nehmen Werte zwischen -1 und +1 an, dürfen aber nicht als Prozentwerte missinterpretiert werden. Die Größe des Koeffizienten ist ein Gradmesser der Stärke des Zusammenhanges. Das Vorzeichen des Koeffizienten beschreibt dessen Richtung: Positiv für einen gleichsinnig proportionalen Zusammenhang (je größer das eine, desto größer das andere), negativ für einen umgekehrt proportionalen Zusammenhang.

## 2.1 DIE ZEHN BEWERTUNGSKATEGORIEN DER SCHREIBAUFGABE

### 2.1.1 DIE ERGEBNISSE DER ZWEITKORREKTUR BERLINWEIT

#### Korrelation von Erst- und Zweitkorrektur: Bleibt die Rangfolge der Arbeiten erhalten?

Tabelle 2.1.1 enthält die Korrelationskoeffizienten, die Stärke (und Richtung) des Zusammenhangs zwischen Erst- und Zweitkorrektur quantifizieren.

**2.1.1 Tabelle:** Die zehn Items der Schreibaufgabe, die Teilpunktsummen zu Inhalt und zur Darstellung sowie die Gesamtpunktzahl für den Teil II. Korrelationen<sup>9</sup> zwischen den Bewertungen aus der Erst- und der Zweitkorrektur.

	Korrelation		Korrelation
I-1: These in eigenen Worten	.26	D-1: Gliederung, Struktur	.25
I-2: Argumente dafür	.36	D-2: Kohärenz	.26
I-3: Argumente dagegen	.36	D-3: Abstrahierende Begriffe	.22
I-4: Gewichtung der Argumente	.34	D-4: Wortwahl	.20
I-5: Eigene Position	.33	D-5: Satzverknüpfung	.24
<i>Inhalt</i>	.49	<i>Darstellung</i>	.43
Teil II	.51		

Die Koeffizienten sind allesamt nicht sonderlich groß. Der Zusammenhang ist positiv, d.h. je besser die Bewertungen bei der einen Korrektur ausfallen, desto besser sind sie auch bei der anderen. Tendenziell - so lässt sich dies interpretieren - wird die Rangfolge der Arbeiten, die sich bei der Erstkorrektur ergab, bei der Zweitkorrektur eingehalten. Allerdings nur in geringem Umfang, wobei die unterschiedlichen Größenordnungen der Korrelationen in den beiden Blöcken zum Teil auf den unterschiedlich großen Wertespektren beruhen: Die Inhaltssitems I-1 ff. können zwischen 0 und 2 schwanken, die Darstellungssitems nur die Werte 0 und 1 annehmen, während dementsprechend 10, 5 und 15 die Maxima der (Teil)Summen sind. Die Möglichkeit größerer Variabilität schlägt sich nahezu zwangsläufig in höheren Korrelationskoeffizienten nieder, die unterschiedlichen Größenordnungen sind demnach teilweise ein statistischer Artefakt.

Gleichwohl ist bedenkenswert, wie niedrig selbst die höchsten der Korrelationen sind. Der Zusammenhang zwischen Erst- und Zweitkorrektur scheint schwach ausgeprägt zu sein.

Darüber hinaus gilt Folgendes: Lügen beispielsweise die Werte der Zweitkorrektur durchgehend um einen Punkt höher als die der Erstkorrektur, betrüge der Korrelationskoeffizient 1, wäre also maximal, und dennoch gäbe es eine systematische und inhaltlich bedeutsame

<sup>9</sup> SPEARMANs  $\rho$ . Alle aufgeführten Korrelationskoeffizienten sind statistisch signifikant größer als 0.

Abweichung. Die weitere Auswertung widmet sich nun direkt diesen Abweichungen, genauer: Sie fragt nach der Größe der Abweichung (Um wie viel liegen die Werte der Erst- und der Zweitkorrektur auseinander?), noch nicht nach ihrer Richtung (Sind der Werte der Zweitkorrektur eher kleiner oder eher größer als die der Erstkorrektur?).

**Absolutbetrag: Wie weit weicht insgesamt die Zweit- von der Erstkorrektur ab?**

**2.1.2 Tabelle:** Die zehn Items der Schreibaufgabe, die Teilpunktsummen zu Inhalt und zur Darstellung sowie die Gesamtpunktzahl für den Teil II. Abweichungen zwischen den Bewertungen aus der Erst- und der Zweitkorrektur. (Grundlage sind die Absolutbeträge der Differenzen; angegeben werden die Mittelwerte über alle N=746 Arbeiten.)

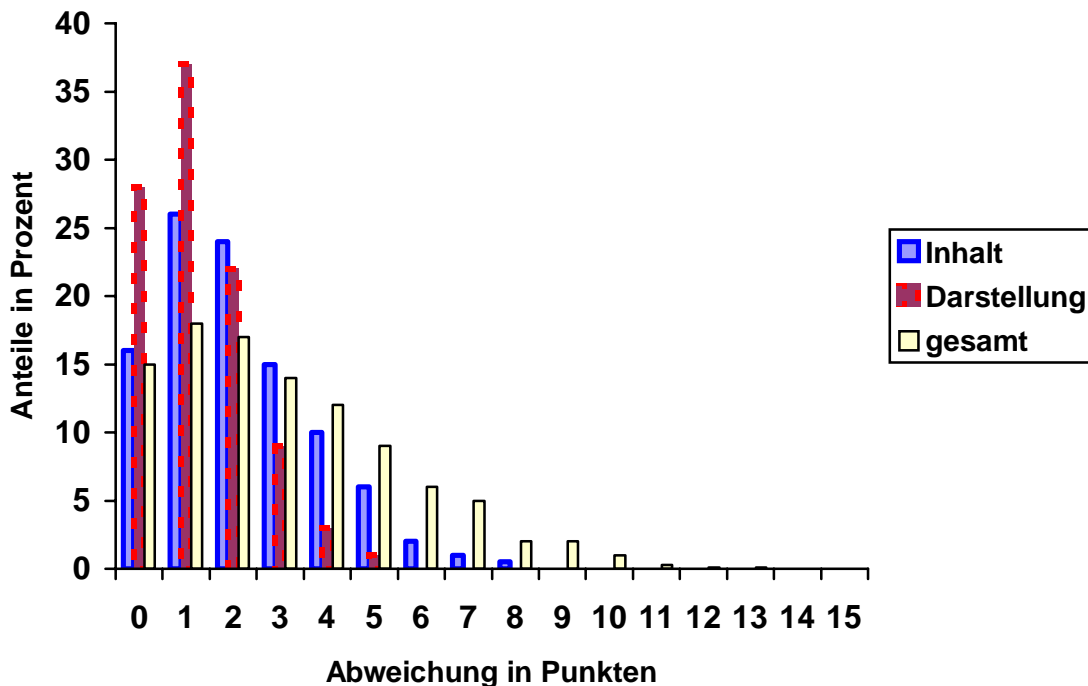
	Abweichung		Abweichung
I-1: These in eigenen Worten	0,7	D-1: Gliederung, Struktur	0,3
I-2: Argumente dafür	0,5	D-2: Kohärenz	0,4
I-3: Argumente dagegen	0,6	D-3: Abstrahierende Begriffe	0,3
I-4: Gewichtung der Argumente	0,6	D-4: Wortwahl	0,4
I-5: Eigene Position	0,6	D-5: Satzverknüpfung	0,4
<i>Inhalt</i>	2,1	<i>Darstellung</i>	1,3
Teil II	3,0		

Bei der Interpretation der Werte sind die unterschiedlichen Wertebereiche zu berücksichtigen: Die möglichen Minima betragen zwar für alle Items und (Teil)Summen 0, die höchstmöglichen jedoch für die Inhaltsitems 2 und somit für die Teilsumme *Inhalt* 10 Punkte, für die Darstellungsitems 1 Punkt, für die Teilsumme *Darstellung* somit 5 und insgesamt für die Schreibaufgabe 15 Punkte. Dies bedeutet beispielsweise für die Items I-1 und D-1: Bei einem Wertebereich von 2 Punkten bzw. 1 Punkt gab es zwischen Erst- und Zweitkorrektur eine mittlere Abweichung von 0,7 bzw. von 0,3. Die Abweichung gemessen am Wertebereich beträgt im ersten Fall also 0,7 von 2, also 35%, im zweiten 0,3 von 1, also 30%. Beide Abweichungen sind demnach größenordnungsmäßig in etwa gleich und angesichts des beengten Wertebereiches recht hoch.

Offensichtlich hängt das Verständnis der vorgegebenen Bewertungsaspekte ziemlich stark von der jeweils korrigierenden Lehrkraft ab. Dies gilt grosso modo für alle Items bei geringfügigen Unterschieden: So scheint I-1 anfälliger für individuelle Interpretationen zu sein als die übrigen Inhaltsitems, während das Bewertungskriterium I-2 hingegen relativ ähnlich bei der Erst- und bei der Zweitkorrektur angelegt wird.

Wie sich die Mittelwerte der Abweichungen von der Erst- zur Zweitkorrektur zusammensetzen, zeigt für die (Teil)Summen die Abbildung 2.1.3. Aufgrund der unterschiedlichen Wertebereiche gilt: Für die Teilsumme *Inhalt* können die Abweichungen zwischen 0 und 10 Punkten schwanken, für die Teilsumme *Darstellung* bis zu 5 Punkten, für die Gesamtsumme kann die Abweichung maximal 15 Punkte erreichen. Abbildung 2.1.3 zeigt, wie häufig die einzelnen möglichen Abweichungen auftreten.

**2.1.3 Abbildung:** Abweichungen zwischen Erst- und Zweitkorrektur (Absolutbeträge). Relative Häufigkeiten für die Teilsummen *Inhalt* und *Darstellung* sowie für die Gesamtsumme der Schreibaufgabe.



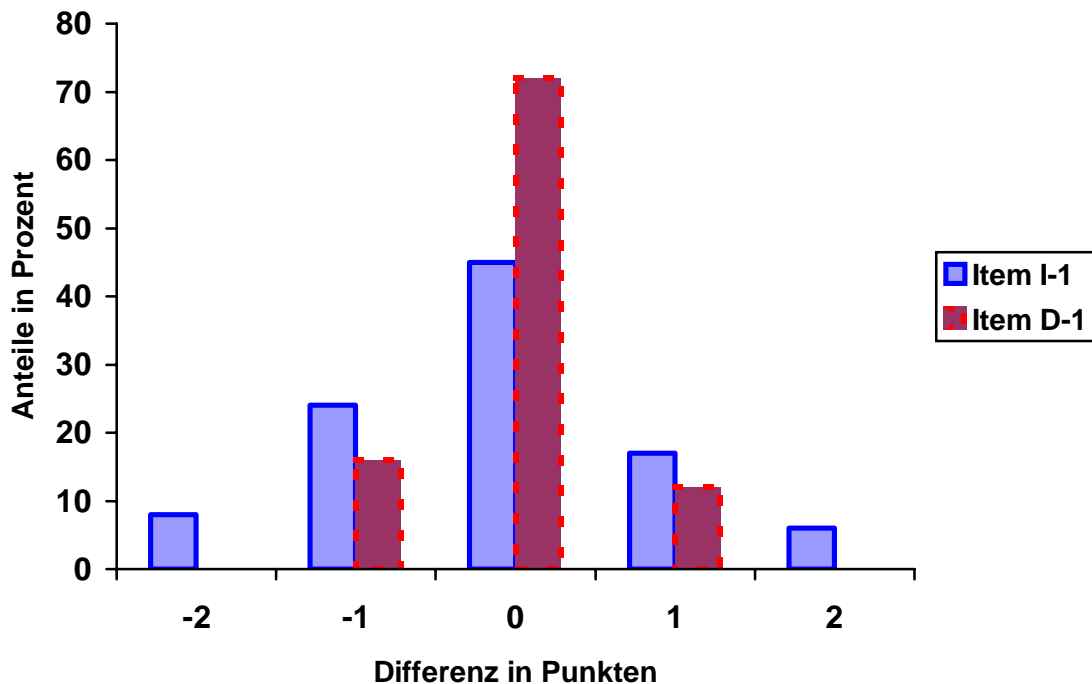
Im Idealfall wäre nur die Kategorie 0 besetzt, wenn nämlich für alle 746 Vergleichsarbeiten Erst- und Zweitkorrektur zur selben Bewertung gekommen wären. Dies ist nicht der Fall. Dennoch ist zu beachten, dass die Kategorie 0 zwar bei keiner der drei Summen die am stärksten vertretene ist, dies aber bereits für die daneben liegende Kategorie der Abweichung um einen Punkt gilt. Abweichungen sind unvermeidlich; tolerieren wir - unter Berücksichtigung der unterschiedlichen Wertebereiche, s.o. - für den Teil *Inhalt* zwei Punkte Abweichung, für den Teil *Darstellung* einen Punkt und für die Gesamtsumme drei Punkte, so liegen in diesen drei Toleranzbereichen 66%, 65% und 64%, also jeweils rund zwei Drittel der Vergleichsarbeiten, ein akzeptabler, wenngleich in Zukunft noch auszubauender Anteil.

Dieser relativ gute Wert war angesichts der recht hohen Abweichungen bei den einzelnen der zehn Bewertungsaspekte nicht zu erwarten. Offensichtlich gleichen sich die Abweichungen bei den einzelnen Items in ihrem Zusammenwirken für die Bewertung der Schreibaufgabe insgesamt wieder aus. Auf diesem Umstand beruht das aus der Testkonstruktion bekannte Phänomen, dass mit wachsender Itemzahl die Zuverlässigkeit des Testgesamtergebnisses steigt. Allerdings wäre es wünschenswert, dass ein ähnlicher oder gar derselbe Punktwert bei ein und derselben Arbeit in der Erst- und in der Zweitkorrektur aus denselben Komponenten sich zusammensetzt, dass also nicht unterschiedliche Gründe zum selben Ergebnis führen. Bereits die einzelnen Bewertungsaspekte müssten so präzise definiert sein, dass sie von allen Lehrkräften in derselben Art und Weise angewandt werden.

### (Gerichtete) Differenz: Führt die Erst- oder die Zweitkorrektur zu besseren Bewertungen?

Der eben diskutierte Umstand wirft die Frage auf, wie die Abweichungen sich zusammensetzen. Neues Element der Analyse ist demnach die Richtung der Abweichung: In wie vielen Fällen liegt die Erst- je über und unter der Zweitkorrektur? Zur Illustration betrachten wir zunächst wieder die beiden Items I-1 und D-1; Abbildung 2.1.4. Maß der Abweichung ist nun die Differenz: Wert der Zweitkorrektur minus Wert der Erstkorrektur ( $K_2 - K_1$ ), also nicht allein der Absolutbetrag. Der mögliche Wertebereich der Differenzen reicht bei den Inhaltsitems I-1 ff. von -2 bis +2, bei den Darstellungsisitem D-1 ff. von -1 bis +1, bei den Teilsummen *Inhalt* von -10 bis +10 und *Darstellung* von -5 bis +5 sowie bei der Gesamtsumme von -15 bis +15.<sup>10</sup> Wünschenswert wäre, dass der Wertebereich nicht ausgeschöpft wird, sondern die meisten Differenzen zwischen Zweit- und Erstkorrektur um die Null herum liegen.

#### 2.1.4 Abbildung: Differenzen "Zweitkorrektur minus Erstkorrektur". Relative Häufigkeiten für die beiden Items I-1 und D-1.



Aus der Abbildung 2.1.4 ergibt sich zweierlei: Zum Einen ist die mittlere Kategorie, jene der Übereinstimmung von Erst- und Zweitkorrektur, am häufigsten vertreten; zum Anderen überwiegen die negativen verglichen mit den entsprechenden positiven Abweichungen (-1 kommt häufiger vor als +1), d.h. es gibt mehr Arbeiten, bei denen die Zweitkorrektur zu einem etwas

<sup>10</sup> Folgender Typ Überlegung führt zu den Wertebereichen: Bei den Inhaltsaspekten können pro Item maximal zwei Punkte vergeben werden. Lautet die Bewertung bei der Zweitkorrektur +2 und bei der Erstkorrektur 0, so beträgt die Differenz +2, im umgekehrten Fall -2 etc.

schlechteren Ergebnis kommt als die Erstkorrektur. Dies gilt für alle Bewertungsaspekte und somit auch für die (Teil)Summen, wie die Tabelle 2.1.5 zeigt: Alle Differenzmittelwerte sind negativ oder Null.<sup>11</sup>

**2.1.5 Tabelle:** Die zehn Items der Schreibaufgabe, die Teilpunktsummen zu Inhalt und zur Darstellung sowie die Gesamtpunktzahl für den Teil II. Differenzen der Bewertungen "Zweitkorrektur minus Erstkorrektur" sowie deren Einzelbewertungen. (Angegeben werden die Mittelwerte über alle N=746 Arbeiten.)<sup>12</sup>

	Differenz	Zweitkorrektur	Erstkorrektur
I-1: These in eigenen Worten	-0,1	0,9	0,8
I-2: Argumente dafür	-0,1	1,1	1,2
I-3: Argumente dagegen	-0,2	1,0	1,2
I-4: Gewichtung der Argumente	-0,2	0,7	0,9
I-5: Eigene Position	-0,1	1,2	1,3
<i>Inhalt</i>	-0,8	4,7	5,5
D-1: Gliederung, Struktur	-0,1	0,7	0,8
D-2: Kohärenz	0,0	0,6	0,6
D-3: Abstrahierende Begriffe	-0,1	0,3	0,4
D-4: Wortwahl	-0,1	0,5	0,6
D-5: Satzverknüpfung	0,0	0,6	0,6
<i>Darstellung</i>	-0,2	2,7	2,9
Teil II	-1,0	7,4	8,4

Die kleinen Mittelwerte, die die Tabelle 2.1.5 verzeichnet, täuschen über das wahre Ausmaß der Abweichungen zwischen Erst- und Zweitkorrektur hinweg, wie Tabelle 2.1.2 zeigt, die auf den Absolutbeträgen beruht. Die Mittelwerte der Tabelle 2.1.5 sind wesentlich kleiner als dort, weil die negativen und die positiven Differenzen sich weitgehend ausgleichen, aber sie enthalten die wichtige Information, dass im Mittel die Erstkorrektur, also jene durch die unterrichtende Lehrkraft, etwas besser ausfällt als die Zweitkorrektur. Genauer, wie es zu den nichtpositiven Mittelwerten kommt, lässt sich der Abbildung 2.1.6 entnehmen, die die Verteilungen der vergebenen Punkte für die Teilsummen wiedergibt. Aus Gründen besserer Darstellbarkeit verzichten wir auf die Wiedergabe der Gesamtsumme, die einen weitaus größeren Wertebereich umfasst als die beiden Teilsummen, sich jedoch aus diesen beiden Komponenten additiv zusammensetzt, so dass das für die beiden Gesagte ebenfalls für die Gesamtbewertung der Schreibaufgabe gilt.

<sup>11</sup> An sich sind das "negative Nullen", die aufgrund der Rundung zu Nullen werden; Beispiel Item D-5:  $K_2=0,58$ ,  $K_1=0,61$ ,  $K_2-K_1=-0,03$ , was zu 0 gerundet wird.

<sup>12</sup> Die Differenzen der Mittelwerte aus der Erst- und der Zweitkorrektur ist statistisch signifikant von 0 verschieden für alle Inhaltsitems sowie für D-3, D-4 und allen Teil- und Gesamtsummen *Inhalt*, *Darstellung*, *Teil II*.

**2.1.6 Abbildung:** Differenzen "Zweitkorrektur minus Erstkorrektur". Relative Häufigkeiten für die Teilsummen *Inhalt* und *Darstellung*.

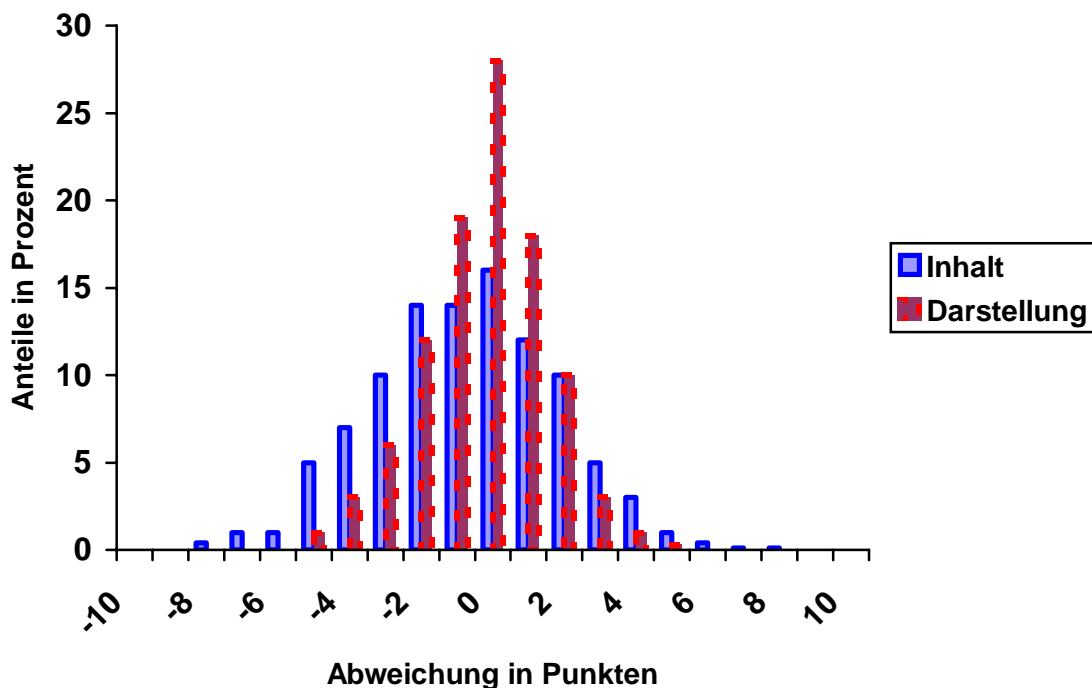


Abbildung 2.1.6 zeigt, dass die mittlere Kategorie jeweils am häufigsten vertreten ist. Sie zeigt auch, dass bei den sich entsprechenden negativen und positiven Kategorien, also z.B. -5 und +5, die negative stärker besetzt ist als die positive, es also mehr Bewertungen gibt, die bei der Erstkorrektur günstiger ausfallen als bei der Zweitkorrektur.

Hintergrund dieser Unterschiede dürften die unterschiedlichen Orientierungsmöglichkeiten der beiden Korrektoren sein: Während bei der Zweitkorrektur nur die Arbeit und die Bewertungskriterien vorliegen, kennt darüber hinaus der Erstkorrektor den/die Schüler/in und wird sich zudem auf den eigenen Unterricht beziehen.

Dennoch bleibt festzuhalten, dass bei einer beträchtlichen Anzahl von Fällen, die erstkorrigierende Lehrkraft strenger urteilt als die zweitkorrigierende. Somit gilt: Es gibt eine Tendenz, dass die unterrichtende Lehrkraft etwas besser bewertet als der Zweitkorrektor, aber das ist nicht die Regel.

---

## 2.1.2 ERGEBNISSE DIFFERENZIERT NACH SCHULARTEN

---

Die schulartpezifische Auswertung beschränkt sich auf die drei aggregierten Variablen der Teilsummen *Inhalt* und *Darstellung* sowie der Gesamtpunktschuldung für die Schreibaufgabe, denn eine Detailanalyse für jeden einzelnen der zehn Bewertungsaspekte soll nicht nur aus Platzgründen vermieden werden, sondern erscheint zudem angesichts der bisherigen Ergebnisse, die keine wesentlichen Unterschiede zwischen den einzelnen Teilkriterien zeigten, nicht erforderlich. Die nachstehend dokumentierten Analysen werden zudem aufgrund der notwendigen Unterteilungen der Stichprobe mit kleineren Fallzahlen auskommen müssen, so dass die zuverlässigeren aggregierten Merkmale geeigneter als die Einzelitems sind.

Wir betrachten zunächst die Korrekturen aller Arbeiten jeweils einer Schulart um dann zu differenzieren: Unterschieden wird, ob die Zweitkorrektur durch eine Lehrkraft derselben Schulart vorgenommen wurde, aus der die zu bewertende Arbeit stammt ( $K1=S=K2$ , also Schulart der Erstkorrektur gleich der Schulart der Zweitkorrektur, und dementsprechend  $K1=S/K2=S$ . Für die einzelnen Schularten gelten sinngemäß die Abkürzungen  $K1=O=K2$ ,  $K1=O/K2\neq O$  etc.).

### Korrelation von Erst- und Zweitkorrektur: Bleibt die Rangfolge der Arbeiten erhalten?

Wir beginnen die Analyse wiederum mit Korrelationen, die den (statistischen) Zusammenhang zwischen der Erst- und Zweitkorrektur im Sinne von "Je größer das Eine, desto größer tendenziell das Andere" quantifizieren.

#### **2.1.7 Tabelle:** Die Teilpunktsummen zu *Inhalt* und zur *Darstellung* sowie die Gesamtpunktzahl für den Teil II. Korrelationen der Bewertungen von Erst- und Zweitkorrektur differenziert nach Schularten. (Vgl. Text.)

Arbeiten	N	<i>Inhalt</i>	<i>Darstellung</i>	<i>Gesamtsumme</i>
<b>Alle Arbeiten</b>	<b>746</b>	<b>.49</b>	<b>.43</b>	<b>.51</b>
aus der <b>Gesamtschule</b>	<b>202</b>	<b>.55</b>	<b>.43</b>	<b>.54</b>
aus der <b>Hauptschule</b>	<b>54</b>	<b>.56</b>	<b>.61</b>	<b>.64</b>
aus der <b>Realschule</b>	<b>172</b>	<b>.41</b>	<b>.40</b>	<b>.46</b>
aus dem <b>Gymnasium</b>	<b>252</b>	<b>.24</b>	<b>.33</b>	<b>.29</b>
aus der <b>Berufsfachschule</b>	<b>66</b>	<b>.48</b>	<b>.16<sup>n.s.</sup></b>	<b>.37</b>

Die Korrelationen sind i.d.R. mittelhoch. Der Koeffizient mit dem Vermerk "n.s." ist nicht signifikant von Null verschieden, könnte "in Wahrheit" also auch Null sein, d.h. dass in diesem Fall Erst- und Zweitkorrektur überhaupt nicht miteinander korreliert wären. Die rechte Spalte *Gesamtsumme* zeigt, dass sich die höchste Korrelation für die Arbeiten aus der Hauptschule, die niedrigste für jene aus dem Gymnasium ergeben haben. Arbeiten aus der Hauptschule scheinen eindeutiger bewertet zu werden als die aus anderen Schularten. Und umgekehrt werden Arbeiten aus dem Gymnasium heterogener bewertet als die der übrigen Schularten.

Dieses erste Ergebnis wird, da Korrelationen Aussagen über die Ähnlichkeit der Rangfolgen machen, anhand der Punktdifferenzen zu überprüfen sein. Doch zunächst werden die Werte der Tabelle 2.1.6 ausdifferenziert, indem wir zwischen schulartidentischen und schulartdifferenten Zweitkorrekturen unterscheiden; vgl. Tabelle 2.1.8.

**2.1.8 Tabelle:** Die Teilpunktsummen zu *Inhalt* und zur *Darstellung* sowie die Gesamtpunktzahl für den Teil II. Korrelationen der Bewertungen von Erst- und Zweitkorrektur differenziert nach Schularten und nach Schulart des Zweitkorrektors. (Vgl. Text.)

	N	<i>Inhalt</i>	<i>Darstellung</i>	<i>Gesamtsumm</i>
<b>K1=S=K2</b>	306	.49	.37	.48
<b>K1=S/K2≠S</b>	440	.48	.48	.53
<b>K1=O=K2</b>	71	.55	.48	.58
<b>K1=O/K2≠O</b>	131	.54	.40	.52
<b>K1=OH=K2</b>	30	.70	.52	.67
<b>K1=OH/K2≠OH</b>	24	.35 <sup>n.s.</sup>	.75	.57
<b>K1=OR=K2</b>	75	.39	.40	.45
<b>K1=OR/K2≠OR</b>	97	.42	.40	.48
<b>K1=OG=K2</b>	92	.08 <sup>n.s.</sup>	.25	.12 <sup>n.s.</sup>
<b>K1=OG/K2≠OG</b>	160	.33	.37	.39
<b>K1=OBF=K2</b>	38	.40	-.02 <sup>n.s.</sup>	.23 <sup>n.s.</sup>
<b>K1=OBF/K2≠OBF</b>	28	.47	.45	.43

Auch hier sind nahezu alle Korrelationen nur mittelhoch. Wir betrachten wiederum zunächst die rechte Spalte, die die Korrelationen des Gesamtergebnisses für die Schreibaufgabe festhält.

In jeder Doppelzeile ist oben die Korrelation vermerkt, wenn beide Korrektoren aus derselben Schulart stammen, darunter, wenn dies nicht der Fall ist. Anzunehmen wäre nun, dass jeweils die obere Korrelation größer ist als die untere. Dies trifft nicht zu, jedenfalls nicht durchgängig. Über die gesamte Stichprobe hinweg (oberste Doppelzeile K1=S) ist mit .48 zu .53 bei Schulartgleichheit die Korrelation kleiner als bei Schulartungleichheit, auch wenn der Unterschied nicht sehr groß und weder statistisch noch inhaltlich bedeutsam ist. Die Vorannahme bestätigt sich nur bei den Gesamt- und bei den Hauptschulen, hingegen nicht bei den Realschulen und insbesondere nicht bei den Gymnasien und Berufsschulen. Für letztere Schularten gilt, dass die Korrelationen (.12 und .23) nicht mehr signifikant von Null verschieden sind, stammt der Zweitkorrektor ebenfalls aus derselben Schulart, und dass zugleich die Differenzen zwischen den Korrelationspaaren (.12/.39 und .23/.43) statistisch ebenfalls nicht signifikant voneinander verschieden sind, es also "in Wahrheit" egal sein könnte, ob die Zweitkorrektur schulartidentisch oder schulartdifferent erfolgt.

Lehrkräfte aus den Gymnasien verweisen darauf, dass die Aufgabenstellung der Vergleichsarbeit 2004 im Vergleich der gängigen Aufsatzpraxis der Schularten von der am Gymnasium üblichen am weitesten abweiche. Die tradierte Arbeitsweise kenne eher die Problemerkörterung mit relativ genauen Anforderungskriterien, so dass nun zwangsläufig Unsicherheiten entstünden, wie die Schreibaufgabe zu bewerten sei, da diese Anforderungen und die Bewertungskriterien zur Vergleichsarbeit nicht deckungsgleich seien.

Die beiden Spalten zu den Teilsummen *Inhalt* und *Darstellung* liefern ein ähnliches Bild. Für beide Komponenten der Bewertung gilt cum grano salis das eben Gesagte, wobei auf den

Sonderfall Hauptschule verwiesen sei. Hier sind die Konstellationen in den Teilbereichen *Inhalt* und *Darstellung* gegensinnig: Die OH-Korrekturen ähneln sich für die inhaltlichen Aspekte mehr, wenn beide Korrektoren aus der Hauptschule kommen, für die darstellerischen Aspekte gilt das Gegenteil, ein Ergebnis, das schulartübergreifend in abgeschwächter Weise ebenfalls zu beobachten ist.

### Unterschied zwischen Erst- und Zweitkorrektur (Absolutbetrag) differenziert nach der Schulart, aus der die Arbeiten stammen

Während Korrelationskoeffizienten quantifizieren, wie weit die Rangfolge von der Erst- zur Zweitkorrektur erhalten blieb, sind die absoluten Abweichungen zwischen den beiden Korrekturen von zentraler Bedeutung für deren numerischer Nähe. Hierüber gibt Tabelle 2.1.9 Auskunft.

**2.1.9 Tabelle:<sup>#</sup> Die Teilpunktsummen zu *Inhalt* und zur *Darstellung* sowie die Gesamtpunktzahl für den Teil II. Unterschiede in den Bewertungen von Erst- und Zweitkorrektur differenziert nach Schularten. (Angegeben werden jeweils die Mittelwerte der Unterschiede zwischen Erst- und Zweitkorrektur in Absolutbeträgen.)<sup>13</sup>**

Arbeiten	N	<i>Inhalt</i>	<i>Darstellung</i>	<i>Gesamtsumme</i>
<b>Alle Arbeiten</b>	<b>746</b>	<b>2,1</b>	<b>1,3</b>	<b>3,0</b>
aus der <b>Gesamtschule</b>	<b>202</b>	<b>1,9</b>	<b>1,2</b>	<b>2,8</b>
aus der <b>Hauptschule</b>	<b>54</b>	<b>1,7</b>	<b>1,0</b>	<b>2,2</b>
aus der <b>Realschule</b>	<b>172</b>	<b>2,2</b>	<b>1,2</b>	<b>3,1</b>
aus dem <b>Gymnasium</b>	<b>252</b>	<b>2,4</b>	<b>1,3</b>	<b>3,3</b>
aus der <b>Berufsfachschule</b>	<b>66</b>	<b>1,8</b>	<b>1,6</b>	<b>2,8</b>

Die rechte Spalte *Gesamtsumme* weist eine Konstellation auf, die bereits aus der Korrelationstabelle 2.1.7 bekannt ist: Die größte Nähe zwischen Erst- und Zweitkorrektur findet sich für die Arbeiten aus der Hauptschule, die geringste bei jenen aus dem Gymnasium.

<sup>#</sup> Die Tabelle A1 aus dem Anhang ergänzt die hier aufgeführten Mittelwerte um die dazugehörigen Streuungen.

<sup>13</sup> Sind die schulartspezifischen Mittelwerte signifikant voneinander verschieden? Varianzanalyse mit Absolutbeträgen: Zum Schätzen der statistischen Signifikanz wird  $\alpha$  halbiert, also für das 5%-Niveau ein  $\alpha$  von 0,025 verwendet. Damit ergibt sich keine Signifikanz für die *Darstellungsmittelwerte*, aber für die beiden anderen von *Inhalt* und *Gesamtsumme*.

Da das Wertespektrum des Bewertungsbereiches *Inhalt* doppelt so groß ist wie desjenigen der *Darstellung*, ist schwierig zu entscheiden, in welchem der Bereiche die größeren Abweichungen auftreten, denn eine reine Verdoppelung ist nicht zulässig, da wir nicht wissen, wie die Korrektoren bei den inhaltlichen Aspekten bewertet hätten, hätte ihnen dort ebenfalls nur die Möglichkeit offen gestanden, keinen oder einen Punkt zu vergeben. Dennoch drängt sich aufgrund des durchgängigen Musters der Eindruck auf, die Kriterien der *Darstellung* seien divergenter verwendet worden als jene des *Inhalts*.

### Unterschied zwischen Erst- und Zweitkorrektur (Differenz) differenziert nach der Schulart, aus der die Arbeiten stammen

Tabelle 2.1.10 ergänzt die bisherigen Ergebnisse durch die hierzu entsprechenden Werte der gerichteten Differenzen.

**2.1.10 Tabelle:<sup>#</sup> Die Teilpunktsummen zu *Inhalt* und zur *Darstellung* sowie die Gesamtpunktzahl für den Teil II. Unterschiede in den Bewertungen von Erst- und Zweitkorrektur differenziert nach Schularten. (Angegeben werden jeweils die Mittelwerte der gerichteten Differenzen von Zweit- minus Erstkorrektur.)<sup>14</sup>**

Arbeiten	N	<i>Inhalt</i>	<i>Darstellung</i>	<i>Gesamtsumme</i>
<b>Alle Arbeiten</b>	<b>746</b>	<b>-0,7</b>	<b>-0,2</b>	<b>-1,0</b>
aus der <b>Gesamtschule</b>	<b>202</b>	<b>-0,4</b>	<b>0,0</b>	<b>-0,4</b>
aus der <b>Hauptschule</b>	<b>54</b>	<b>-0,3</b>	<b>0,5</b>	<b>0,2</b>
aus der <b>Realschule</b>	<b>172</b>	<b>-0,7</b>	<b>-0,3</b>	<b>-1,1</b>
aus dem <b>Gymnasium</b>	<b>252</b>	<b>-1,2</b>	<b>-0,5</b>	<b>-1,8</b>
aus der <b>Berufsfachschule</b>	<b>66</b>	<b>-0,1</b>	<b>-0,1</b>	<b>-0,2</b>

Zunächst ergibt sich - und dies nahezu durchgängig - das bereits aus dem Abschnitt 2.1.1 bekannte Bild negativer Differenzen. Mit einer Ausnahme gibt es in allen Schularten die Tendenz, bei der Zweitkorrektur etwas strenger zu bewerten als bei der Erstkorrektur. Und es gilt zugleich das ebenfalls bereits Festgestellte: Eine Tendenz ist keine Regel, d.h. die Mittelwerte dürfen nicht vergessen machen, dass es ebenfalls viele Zweitkorrekturen gibt, die zu besseren Ergebnissen als die Erstkorrekturen führen.

<sup>#</sup> Vgl. die ergänzende Tabelle A1 im Anhang, die die Streuungen angibt. Die dortige Tabelle A2 zeigt die Ausgangsgrößen, die zu den hier aufgelisteten Differenzen führen.

<sup>14</sup> Sind die schulartspezifischen Mittelwerte signifikant voneinander verschieden? Varianzanalyse mit den Differenzen: Zum Schätzen der statistischen Signifikanz wird  $\alpha$  halbiert, also für das 5%-Niveau ein  $\alpha$  von 0,025 verwendet. Alle drei Typen von Mittelwerten, *Darstellung*, *Inhalt* und *Gesamtsumme*, scheinen demnach statistisch signifikant zu sein.

Bemerkenswert die Hauptschule als Ausnahme: Die Zweitkorrektoren bewerten milder, allerdings nur im Bewertungsbereich *Darstellung*. Im Bereich *Inhalt* hingegen folgt die Zweitkorrektur dem Muster der anderen Schularten, nämlich dem einer tendenziell strengeren Bewertung.

### **Unterschiede zwischen Erst- und Zweitkorrektur (Absolutbeträge und Differenzen) zusätzlich differenziert nach der Schulart der Zweitkorrektur**

Im nächsten Schritt wird die Tabelle 2.1.9 weiter differenziert, indem unterschieden wird, ob die Zweitkorrektur schulartidentisch oder schulartdifferent erfolgte; vgl. Tabelle 2.1.11. Anzunehmen wäre, dass die schulartidentische Korrektur zu geringeren Abweichungen zwischen der Erst- und der Zweitkorrektur als die schulartdifferente führt.

In Tabelle 2.1.11 müssten in jeder Doppelzeile die oberen Werte kleiner sein als die unteren, was aber nicht durchgängig der Fall ist: Stimmgig mit der Vorannahme sind die Werte für die Gesamt- und die Realschule. Trotz der Unsicherheit, die mit den geringen Fallzahlen bei Haupt- und Berufsschule verbunden ist, ist dennoch bemerkenswert, wie die Konstellationen in den beiden Teilbereichen gegensinnig sind, nämlich hypothesenkonform für *Inhalt*, während für die darstellerischen Aspekte die Abweichungen zwischen Erst- und Zweitkorrektur größer sind bei den schulidentischen Korrekturen. Dieses hypothesenwidrige Resultat gilt (und hier bei hohen Fallzahlen) generell für das Gymnasium.

Vor einer weiteren Analyse ergänzen wir diese Ergebnisse durch die Differenzen zwischen Erst- und Zweitkorrektur, berücksichtigen also die Richtung der Abweichungen; vgl. Tabelle 2.1.12.<sup>15</sup>

Die obersten Doppelzeilen in den Tabellen 2.1.11 und 2.1.12 legen nahe anzunehmen, das Ausmaß der Unterschiede von der Erst- zur Zweitkorrektur sei unabhängig davon, ob die Zweitkorrektur aus derselben oder einer anderen Schulart stamme. Mit Abweichungen von 2,0 und 2,2 bzw. von -0,7 und -0,8 beim *Inhalt* sowie von 1,3 und 1,2 bzw. -0,3 und -0,2 bei der *Darstellung* erhalten wir global gesehen jeweils dieselben Größenordnungen. Aber das, was über alle Schularten hinweg gilt, gestaltet sich - wie bereits angedeutet - im Einzelnen sehr unterschiedlich.

---

<sup>15</sup> Versuchen wir wie bei den Gegebenheiten der Tabellen 2.1.9 und 2.1.10 Schätzungen für die statistische Signifikanz durch Halbierung von  $\alpha$  zu erhalten, so ergibt sich in beiden Tabellen 2.1.11 und 2.1.12 nur in letzterer eine signifikante Paarung, nämlich die gymnasialen Differenzmittelwerte des Bereiches *Darstellung*.

**2.1.11 Tabelle:<sup>#</sup>** Die Teilpunktsummen zu *Inhalt* und zur *Darstellung* sowie die Gesamtpunktzahl für den Teil II. Abweichungen in den Bewertungen von Erst- und Zweitkorrektur differenziert nach Schularten. (Angegeben werden jeweils die Mittelwerte der Absolutbeträge der Unterschiede zwischen Erst- und Zweitkorrektur.)

	N	<i>Inhalt</i>	<i>Darstellung</i>	<i>Gesamtsumm</i>
K1=S=K2	306	2,0	1,3	3,0
K1=S/K2≠S	440	2,2	1,2	3,0
K1=O=K2	71	1,8	1,1	2,5
K1=O/K2≠O	131	2,0	1,3	2,9
K1=OH=K2	30	1,4	1,2	2,1
K1=OH/K2≠OH	24	2,2	0,8	2,3
K1=OR=K2	75	2,0	1,2	2,9
K1=OR/K2≠OR	97	2,4	1,3	3,3
K1=OG=K2	92	2,6	1,5	3,7
K1=OG/K2≠OG	160	2,3	1,1	3,1
K1=OBF=K2	38	1,7	1,8	2,8
K1=OBF/K2≠OBF	28	1,9	1,3	2,8

**2.1.12 Tabelle:<sup>#</sup>** Die Teilpunktsummen zu *Inhalt* und zur *Darstellung* sowie die Gesamtpunktzahl für den Teil II. Unterschiede in den Bewertungen von Erst- und Zweitkorrektur differenziert nach Schularten. (Angegeben werden jeweils die Mittelwerte der Differenzen zwischen Erst- und Zweitkorrektur.)

	N	<i>Inhalt</i>	<i>Darstellung</i>	<i>Gesamtsumm</i>
K1=S=K2	306	-0,7	-0,3	-1,0
K1=S/K2≠S	440	-0,8	-0,2	-0,9
K1=O=K2	71	-0,6	-0,2	-0,8
K1=O/K2≠O	131	-0,3	0,1	-0,2
K1=OH=K2	30	0	0,4	0,5
K1=OH/K2≠OH	24	-0,8	0,6	-0,2
K1=OR=K2	75	-0,4	-0,2	-0,6
K1=OR/K2≠OR	97	-1,0	-0,4	-1,5
K1=OG=K2	92	-1,4	-0,9	-2,3
K1=OG/K2≠OG	160	-1,1	-0,4	-1,4
K1=OBF=K2	38	-0,1	-0,2	-0,3
K1=OBF/K2≠OBF	28	-0,2	0,0	-0,1

<sup>#</sup> Vgl. in Ergänzung hierzu die Tabelle A1 und A2 aus dem Anhang.

Unter Bezug auf die Leitfragen aus dem Abschnitt 1 seien die wesentlichen Ergebnisse tabellarisch zusammengefasst.<sup>16</sup>

**1. Sind die Bewertungen von Erst- und Zweitkorrektur ähnlicher, wenn der Zweitkorrektor aus derselben Schulart kommt als aus einer anderen?**

	<u>Inhalt</u>	<u>Darstellung</u>	<u>Teil II insgesamt</u>
<b>Insgesamt</b>	(ja) <sup>17</sup>	---- <sup>18</sup>	----
O	(ja)	(ja)	ja
OH	ja	nein	(ja)
OR	ja	----	ja
OG	nein	nein	nein
OBF	(ja)	nein	----

Entgegen der Vorannahme fallen Ergebnisse im Bewertungsbereich *Darstellung* aus: Hier häufen sich die Fälle, bei denen die Zweitkorrekturen zu größeren Abweichungen bei schulartidentischen als bei schulartdifferenten Korrektoren führen. Dabei ist die Größenordnung der Abweichungen angesichts des kleineren Wertespektrums als beim Bewertungsbereich *Inhalt* bemerkenswert.

Unter den Schularten weisen - wie bereits erwähnt - neben dem Gymnasium auch die Haupt- und die Berufsschulen eine hypothesenwidrige Konstellation auf, wobei bei letzteren beiden allerdings die geringen Fallzahlen zur Vorsicht bei der Interpretation mahnen.

**2. Gibt es schulartspezifisch größere oder kleinere Differenzen zwischen der Erst- und der Zweitkorrektur?**

Die höchsten Abweichungen finden wir bei den Arbeiten aus den Gymnasien und der Realschule, die geringsten bei jenen aus der Hauptschule; vgl. Tabelle 2.1.9.

Die Ergebnisse weisen auf Besonderheiten im Korrekturverhalten der Lehrkräfte aus den Gymnasien hin. Dies dürfte mit den immer wieder berichteten schulspezifischen Kommunikationsverhalten zusammenhängen. Allem Anschein nach gibt es in den meisten Gymnasien keine Tradition der systematischen und kontinuierlichen Diskussion zentraler Themen wie Inhalt und Gestaltung des Unterrichts im Gegensatz z.B. zu den Gesamtschulen. Bei der kriterialen Bewertung sind sich Studenräte/innen unsicher; es gibt keine expliziten Absprachen, sondern nur "stille Übereinkünfte".

<sup>16</sup> **Hinweis.** Welche Möglichkeiten einer Feinanalyse sich ergäben, wären die Fallzahlen größer, wird im Anhang A3 exemplarisch aufgezeigt.

<sup>17</sup> (ja): Eingeschränktes Ja.

<sup>18</sup> ----: Kein Unterschied.

## 2.2 DIE FORMALFEHLER

Die Datenbasis für die Auswertung der Formalfehler ist wesentlich schmaler als bei den bisher behandelten Bewertungsaspekten. Es gibt insgesamt 271 Arbeiten, von denen sowohl für die Erst- wie für die Zweitkorrektur die Gesamtzahl der Wörter angegeben wurden, und 283 Arbeiten mit beiden Angaben für die drei Fehlertypen (Orthografie, Grammatik, Interpunktion). Die Werte des Fehlerquotienten liegen nur von den 233 Arbeiten vor, von denen wir sowohl die Gesamtzahl an Wörtern als die Gesamtzahl der Fehler kennen. Wie sich diese Arbeiten auf die Schularten verteilen, hält Tabelle 2.2.1 fest.

**2.2.1 Tabelle:** Schulart, aus der die Vergleichsarbeiten Deutsch stammen, und ihre Aufteilung zur Zweitkorrektur auf die Schularten. (Angegeben sind die absoluten Häufigkeiten und die Zeilenprozente; in der Spalte *gesamt* zusätzlich die Spaltenprozente.)

Erstkorrektur	Schulart der Zweitkorrektur					gesamt	
	O	OH	OR	OG	OBF		
O	30 43%	7 10%	10 14%	15 21%	8 11%	70 100%	30%
OH	0 0%	10 56%	0 0%	4 22%	4 22%	18 100%	8%
OR	2 3%	6 10%	35 56%	13 21%	6 10%	62 100%	27%
OG	2 3%	20 26%	17 22%	32 42%	6 8%	77 100%	33%
OBF	0 0%	0 0%	0 0%	5 83%	1 17%	6 100%	3%
gesamt	34 15%	43 18%	62 27%	69 30%	25 11%	233 100%	100%

Tabelle 2.2.1 macht deutlich, dass aufgrund der wesentlich niedrigeren Fallzahlen die nachstehende Auswertung der Formalfehler größeren Einschränkungen unterliegt als die bisherige Auswertung der Bewertungsaspekte *Inhalt* und *Darstellung*. Wir vervollständigen die Aufstellung zur schulartspezifischen Verteilung der Arbeiten aus dem Abschnitt 1.1.

Aufteilung bei der	Grundgesamtheit	Stichprobe Deutsch insgesamt	Zweitkorrektur Aufgabe 18	Zweitkorrektur Formalfehler
O	26%	27%	27%	30%
OH	10%	8%	7%	8%
OR	20%	21%	23%	27%
OG	31%	32%	34%	33%
OBF	13%	12%	9%	3%.

Zwar liegen wesentlich weniger Arbeiten vor, in denen Angaben zu den Formalfehlern als zu den Bewertungsaspekten der Aufgabe 18 gemacht werden, aber die Auswahl dieser Arbeiten weist nur teilweise wesentliche Abweichungen zwischen den beiden Bewertungsbereichen oder zur Grundgesamtheit hin auf: Am gravierendsten ist der nahezu völlige Ausfall von auswertbaren Angaben aus der Berufsschule. Ansonsten wiederholt sich die leichte Unterrepräsentanz der Hauptschule und die geringfügige Überrepräsentanz des Gymnasiums, die sich bei der Realschule deutlich vergrößert, während sie bei der Gesamtschule nur in diesem Bewertungsbereich auftritt. Hinsichtlich des Merkmals Schulart hat eine substanzielle systematische Auswahl bei der Reduktion der 746 auf die knapp 233 Arbeiten nur hinsichtlich Berufs- und Realschule - und dies gegenläufig - stattgefunden.

---

## 2.2.1 DIE ERGEBNISSE DER ZWEITKORREKTUR BERLINWEIT

---

### Korrelation von Erst- und Zweitkorrektur: Bleibt die Rangfolge der Arbeiten erhalten?

Tabelle 2.2.2 enthält die Korrelationskoeffizienten, die Stärke (und Richtung) des Zusammenhangs zwischen Erst- und Zweitkorrektur quantifizieren.

### **2.2.2 Tabelle:** Die Formalfehler (Fehler der sprachlichen Richtigkeit). Korrelationen<sup>19</sup> zwischen den Bewertungen aus der Erst- und der Zweitkorrektur.

	Korrelation
F-1: Wortanzahl	.68
F-2: Orthographiefehler	.63
F-3: Grammatikfehler	.35
F-4: Interpunktionsfehler	.65
F-5: Gesamtfehlerzahl	.72
F-6: Fehlerquotient <sup>20</sup>	.77

Die Korrelationskoeffizienten sind zwar deutlich größer als die der Bewertungsisems I-1 ff. und D-1 ff. (vgl. Tabelle 2.1.1), wesentlich höhere wären aber zu erwarten gewesen, denn abgesehen von immer auftretenden Ungenauigkeiten beim Zählen müssten sich bei den Deutscharbeiten immer dieselben Fehleranzahlen ergeben, da es sich um Fehlertypen handelt, die durch einheitliche und allgemein bekannten Regeln definiert und im Prinzip objektivierbar sind.

---

<sup>19</sup> SPEARMANs  $\rho$

<sup>20</sup> Da in den Fehlerquotienten zwei Zufallsvariable eingehen, zwei Merkmale, die zufälligen Schwankungen unterliegen, und dies in der Konstellation Quotient, ist der Korrelationskoeffizient als eine rein heuristische Größe zu betrachten.

Die Rangfolge der Koeffizienten ist plausibel: Der niedrigste Wert ergibt sich für die Grammatikfehler, über die am schwersten Einigkeit zu erzielen werden mag, während der Fehlerquotient die höchste Korrelation aufweist, also eine aus mehreren Komponenten zusammengesetzte Größe, von denen wir wissen, dass sie tendenziell eine größere Zuverlässigkeit besitzt als ein Einzelmerkmal; vgl. weiter oben den Kommentar zur Abbildung 2.1.3.

**Absolutbetrag: Wie weit weicht insgesamt die Zweit- von der Erstkorrektur ab?**  
**Differenz: Führt die Erst- oder die Zweitkorrektur zu besseren Bewertungen?**

Wie steht es um die absoluten Abweichungen und den Differenzen zwischen den Werten der Erst- und der Zweitkorrektur? Antwort hierauf liefert die Tabelle 2.2.3 (die Analoga zu den Tabellen 2.1.2 und 2.1.5).

**2.2.3 Tabelle: Die Formalfehler (Fehler der sprachlichen Richtigkeit). Unterschiede in den Bewertungen aus der Erst- und der Zweitkorrektur.** (Grundlage sind die Absolutbeträge der Differenzen (Abweichung) und die gerichteten Differenzen  $K_2 - K_1$  (Zweit- minus Erstkorrektur); angegeben werden die Mittelwerte über die  $N=233$  Arbeiten.)

	Abweichung	Differenzen
F-1: Wortanzahl	66	-53
F-2: Orthographiefehler	3,6	-0,8
F-3: Grammatikfehler	2,9	-0,1
F-4: Interpunktionsfehler	3,6	-0,8
F-5: Gesamtfehlerzahl	7,7	-1,6
F-6: Fehlerquotient	2,4	0,5

Aus wie viel Wörtern der Text besteht, den Schüler/innen beim Bewältigen von Schreibaufgaben produzieren, wird i.d.R. geschätzt, womit eine gewisse Unsicherheit verbunden ist, die sich in der Tabelle 2.2.3 niederschlägt. Im Mittel beträgt die Abweichung zwischen der Zählung aus der Zweit- und aus der Erstkorrektur 66 Wörter, wenn außen vor bleibt, ob die Abweichung nach unten oder nach oben erfolgt. Berücksichtigen wir diese Richtung, so dass positive und negative Abweichungen gegeneinander verrechnet werden, dann ergibt sich ein Saldo im Schnitt von -53 Wörtern, d.h. tendenziell führt die Zweitkorrektur zu niedrigeren Wortumfängen als die Erstkorrektur. In 36% der Fälle liegen die Zweitschätzungen unter denen der Erstschätzungen; aber bei immerhin rund der Hälfte der Arbeiten kommen die Schätzungen zu denselben Ergebnissen. Dennoch können die Abweichungen zwischen der Erst- und der Zweitkorrektur enorm sein: Beziehen wir die Zweit- auf die Erstangabe, so schwanken diese Anteile von 24% bis 350%!

Die absoluten Abweichungen bei den einzelnen Fehlerarten sind nicht sonderlich groß, dennoch bemerkenswert. Von besonderem Interesse ist der Vergleich der Werte mit den Korrelationskoeffizienten aus Tabelle 2.2.2, aufgrund dessen höhere Abweichungen zwischen den Ergebnissen aus Erst- und Zweitkorrektur hätten erwartet werden können. Zu beachten ist

allerdings ein Unterschied dessen, was in den beiden Tabellen quantifiziert wird: Die Korrelationen in Tabelle 2.2.2 messen, wie weit die Anordnung der Arbeiten bei beiden Korrekturen dieselbe ist (Wie weit sind die besseren Arbeiten der Erst- auch die besseren Arbeiten der Zweitkorrektur?), während die Mittelwerte die Unterschiede zwischen den beiden Korrekturen zur Grundlage haben. Diese Unterschiede, auch wenn sie nicht sehr groß sind, verändern offensichtlich die Rangfolge der Arbeiten von der Erst- zur Zweitkorrektur, was zu den niedrigen Korrelationen führt. Die Sachlage wird durch die Tabelle 2.2.4 erhellt, die gewissermaßen hinter die Mittelwerte schaut, indem sie die Differenzen bei jeder einzelnen Arbeit in drei Kategorien einteilt: Erstkorrektur ergab größere Werte als die Zweitkorrektur oder umgekehrt oder beide Korrekturen kamen zum selben Ergebnis.

**2.2.4 Tabelle: Die Formalfehler (Fehler der sprachlichen Richtigkeit). Unterschiede zwischen den Bewertungen aus der Erst- und der Zweitkorrektur.** (Grundlage sind die gerichteten Differenzen  $K2 - K1$  (Zweit- minus Erstkorrektur);  $N=233$  Arbeiten. Angegeben wird jeweils, wie hoch die Anteile von Arbeiten sind, bei denen die Erst- oder die Zweitkorrektur zu höheren Werten kam oder zum selben Ergebnis.)<sup>21</sup>

	K2 < K1 K2 << K1	K2 = K1 K2 $\cong$ K1	K2 > K1 K2 >> K1
F-1: Wortanzahl	36% 27%	49% 63%	15% 10%
F-2: Orthographiefehler	42% 22%	17% 61%	41% 17%
F-3: Grammatikfehler	43% 21%	16% 58%	41% 21%
F-4: Interpunktionsfehler	42% 24%	17% 55%	41% 22%
F-5: Gesamtfehlerzahl	45% 25%	9% 54%	46% 21%
F-6: Fehlerquotient	28% 12%	19% 68%	53% 20%

Bei allen Fehlertypen und bei der Gesamtzahl an Worten ergibt die Zweitkorrektur häufiger niedrigere Werte als die Erstkorrektur, daher die negativen Differenzen in der Tabelle 2.2.3. Beim Fehlerquotienten hingegen ist die Differenz positiv, d.h. die Zweitkorrektur ermittelt einen höheren Anteil an Fehlern als die Erstkorrektur. Der Grund ist in der Bildung des Quotienten *Gesamtfehlerzahl/Wortanzahl* zu suchen: Die Zweitkorrektur kommt beim Wortumfang in stärkerem Ausmaß zu niedrigeren Werten als bei den Fehlertypen, so dass im Vergleich zur Erstkorrektur der Anteil der Fehler sich erhöht.

<sup>21</sup> Da sich Differenzen zwischen den Werten der Erst- und der Zweitkorrektur nie werden vermeiden lassen, machen wir jeweils zwei Angaben: Zum Einen wird die Aufteilung zugrundegelegt, die auf der exakten Gleichheit beruht ( $K2=K1$ ); zum Anderen tolerieren wir folgende Abweichungen in der Mittelkategorie, die dementsprechend durch  $K2\cong K1$  (ungefähre Gleichheit) gekennzeichnet wird: F-1:  $\pm 10$ , F-2 bis F-4:  $\pm 2$ , F-5:  $\pm 4$ , F-6:  $\pm 2$ .

Während also bei den Bewertungsaspekten *Inhalt* und *Darstellung* eine Tendenz besteht, dass die Zweitkorrektur zu schlechteren Bewertungen als die Erstkorrektur führt, ist es bei den Formalfehlern mit Ausnahme des daraus abgeleiteten Fehlerquotienten genau umgekehrt. Zu unterstreichen ist, dass es sich um eine Tendenz, keine durchgängige Regel handelt, wie die Tabelle 2.2.4 deutlich macht: Bei den drei Fehlertypen sind die positiven und die negativen Differenzen in etwa gleich häufig vertreten.

Bemerkenswert ist aber, wie relativ selten bei diesen objektivierbaren Fehlern Erst- und Zweitkorrektur zum selben Ergebnis kommen.

## 2.2.2 ERGEBNISSE DIFFERENZIERT NACH SCHULARTEN

Die schulartspezifische Auswertung weist aufgrund der niedrigen Fallzahlen die Schularten Haupt- und Berufsfachschule nicht gesondert aus. Ihre Werte gehen aber immer dort mit ein, wo *Alle Arbeiten* ö.Ä. steht. Da die nachstehenden Analysen mehr hypothesenspendenden denn hypothesenprüfenden Charakter haben, beschränkt sich die Darstellung auf die drei Merkmale Wortanzahl, Gesamtfehlerzahl und Fehlerquotient. Wir betrachten zunächst die Korrekturen aller Arbeiten jeweils einer Schulart, um dann zu differenzieren. Es gilt die im Abschnitt 2.1.2 eingeführte Notation. Begonnen wird wiederum mit den Korrelationen zwischen Erst- und Zweitkorrektur.

**2.2.6 Tabelle:** Die Formalfehler (Fehler der sprachlichen Richtigkeit). Korrelationen<sup>22</sup> zwischen den Bewertungen aus der Erst- und der Zweitkorrektur differenziert nach den Schularten, aus denen die Arbeiten stammen.

Arbeiten	N	Korrelation von Erst- und Zweitkorrektur		
		Wortanzahl	Gesamtfehlerzahl	Fehlerquotient
Alle Arbeiten	233	.68	.72	.77
K1=S=K2	108	.57	.65	.74
K1=S/K2≠S	125	.74	.77	.80
Gesamtschule	70	.52	.68	.80
K1=O=K2	30	.32 <sup>n.s.</sup>	.54	.84
K1=O/K2≠O	40	.69	.76	.79
Realschule	62	.84	.74	.75
K1=OR=K2	35	.74	.76	.75
K1=OR/K2≠OR	27	.95	.76	.75
Gymnasium	77	.62	.66	.66
K1=OG=K2	32	.66	.70	.64
K1=OG/K2≠OG	45	.57	.62	.45

<sup>22</sup> SPEARMANs  $\rho$

Die Korrelationskoeffizienten sind zwar alle mittelhoch, aber für an sich objektiv zu ermittelnden Größen sehr niedrig.

Höhere Korrelationen wären bei schulidentischen als bei schuldifferenten Korrekturen zu erwarten. In den meisten Fällen trifft das Gegenteil zu; im Hinblick auf die Ergebnisse aus dem Abschnitt 2.1.2 ist die gymnasiale Ausnahme mit Werten, die der Vorannahme entsprechen, bemerkenswert.

Tabelle 2.2.7 thematisiert die Abweichungen von der Erst- und zur Zweitkorrektur.

**2.2.7 Tabelle:** Die Formalfehler (Fehler der sprachlichen Richtigkeit). Unterschiede in den Bewertungen aus der Erst- und der Zweitkorrektur differenziert nach Herkunftsschulart der Erst- und Zweitkorrektoren. (Grundlage sind die Absolutbeträge der Differenzen (Abweichung; *ABS*) und die gerichteten Differenzen  $K2 - K1$  (Zweit- minus Erstkorrektur; *Diff*); angegeben werden die Mittelwerte.)

Arbeiten	N	Unterschiede zwischen Erst- und Zweitkorrektur					
		Wortanzahl		Gesamtfehlerzahl		Fehlerquotient	
		ABS	Diff	ABS	Diff	ABS	Diff
Alle Arbeiten	233	66	-53	7,7	-1,6	2,4	0,5
K1=S=K2	108	67	-55	8,3	-2,0	2,6	0,7
K1=S/K2≠S	125	65	-51	7,1	-1,3	2,2	0,3
Gesamtschule	70	106	-85	11,7	-5,2	2,8	1,1
K1=O=K2	30	132	-107	13,5	-4,3	3,3	2,5
K1=O/K2≠O	40	87	-69	10,4	-5,9	2,5	0,1
Realschule	62	29	-24	6,7	2,5	2,2	1,2
K1=OR=K2	35	35	-31	6,1	0,3	1,7	0,9
K1=OR/K2≠OR	27	20	-17	7,4	5,3	2,7	1,5
Gymnasium	77	78	-62	4,8	-0,8	1,6	0,3
K1=OG=K2	32	63	-50	5,1	0,0	1,6	0,5
K1=OG/K2≠OG	45	88	-71	4,6	-1,3	1,6	0,2

Die schulspezifischen Zeilen zeigen, dass hinter den oberen globalen Werte, die über alle Schularten hinweg ermittelt wurden, stark unterschiedliche Verhältnisse im Einzelnen stehen. Während es z.B. beim Merkmal *Wortanzahl* insgesamt keinen Unterschied macht, ob die Zweitkorrektur schulartgleich oder schulartdifferent erfolgte (die Abweichungen betragen 67 bzw. 65 Wörter), so sind etwa bei der Gesamtschule die Abweichungen wesentlich größer und dies zumal, wenn der Zweitkorrektor ebenfalls aus der Gesamtschule kommt. Bei diesem Merkmal zeigen die negativen Vorzeichen in der Spalte *Diff*, dass im Schnitt die Zweit-

korrektur zu geringeren Wortumfängen gelangt als die Erstkorrektur. Aufgrund der Abweichungen bei der Gesamtfehlerzahl ist der Fehlerquotient wiederum durchweg positiv, d.h. die Zweitkorrektur kommt hier - wie bei den inhaltlichen und darstellerischen Bewertungskriterien des Abschnitts 2.1 - zu tendenziell schlechteren Ergebnissen.

**Insgesamt gilt zweierlei:** Es lässt sich kein einheitliches Muster über die Schularten hinweg erkennen - was an den geringen Fallzahlen und der mangelnden Repräsentativität liegen mag -- und die Abweichungen sind für im Prinzip objektive Merkmale erstaunlich hoch und somit sehr unzuverlässige Größen.

---

### **3 RESÜMEE: ZUSAMMENFASSUNG UND SCHLUSSFOLGERUNGEN**

---

#### **3.1 ZUSAMMENFASSUNG**

---

##### **A AUSGANGSLAGE**

---

###### **(1) Relevanz der Untersuchung**

Ihre zentrale Funktion können Vergleichsarbeiten nur dann erfüllen, wenn ihre Ergebnisse tatsächlich vergleichbar sind, wenn also insbesondere die Auswertungsobjektivität gesichert ist. Eine Maßnahme in diesem Sinne ist die Standardisierung der Arbeiten, die weitgehend aus multiple-choice-Aufgaben bestehen. Eine weitere Maßnahme sind die Auswertungshinweise, die präzise und umfassend sein müssen. Besonders schwer ist Auswertungsobjektivität bei Schreibaufgaben zu erreichen. Im Fach Deutsch ist der zweite Teil der Vergleichsarbeit 2004 eine derartige Schreibaufgabe gewesen und trug mit 15 von insgesamt 50 Punkten erheblich zur Gesamtbewertung bei. Wie stabil die bei der Erstkorrektur vergebenen Punkte sind, sollte durch eine Zweitkorrektur überprüft werden.

###### **(2) Fragestellung**

- \* Ist bei den Vergleichsarbeiten Klasse 10 (im Fach Deutsch) die Auswertungsobjektivität so weit gesichert, dass tatsächlich vergleichbare Bewertungen an den Berliner Schulen erfolgen und somit zuverlässige Bezugswerte ermittelt werden können?
- \* Treten schulartspezifische Unterschiede auf:
  - Sind die Bewertungen ähnlicher, wenn der Zweitkorrektor aus derselben Schulart als aus einer anderen stammt?
  - Gibt es schulartspezifisch größere oder kleinere Differenzen zwischen der Erst- und der Zweitkorrektur, auch wenn beide Korrektoren derselben Schulart angehören?
- \* Kann die Praxis beibehalten werden, die Arbeiten von den Lehrkräften korrigieren zu lassen, die die betroffenen Klassen unterrichten?
- \* Wie weit müssen die bisherigen Auswertungshinweise und Bewertungskategorien differenziert und präzisiert werden?

###### **(3) Anlage und Durchführung der Untersuchung**

Im Frühjahr 2004, als die Vergleichsarbeiten geschrieben wurden, war eine Zufallsstichprobe von 80 Klassen (und somit auch 80 Schulen) gezogen worden, die die Ergebnisse zurückmelden sollten. Im Fach Deutsch erging darüber hinaus die Bitte, zugleich einen halben Klassensatz an Deutscharbeiten mitzuschicken, auf denen keinerlei Eintragungen vorzunehmen waren. Die Erstkorrektur musste also auf einer Kopie durchgeführt werden oder eine unkorrigierte Kopie war einzusenden. Diese Arbeiten wurden gleichmäßig auf (fast) alle Nichtprobenschulen verteilt mit der Bitte um Korrektur, die in Unkenntnis des Erstergebnisses und unbeeinflusst von irgendwelchen Anmerkungen erfolgte. Die Herkunft der Arbeiten war - so weit erforderlich - unkenntlich gemacht worden. Beim Versand wurde darauf geachtet, die Arbeiten jeweils einer Schulart ungezielt auf alle Schularten zu verteilen.

Grundlage der Auswertung bildeten 746 Arbeiten, die sich bei geringfügigen Abweichungen repräsentativ über die Schularten verteilen; vgl. hierzu Tabelle 1.2.

#### (4) Merkmale, die der Analyse zugrundeliegen

Im Mittelpunkt der Auswertung stehen zwei Blöcke von Merkmalen. Zum einen die zehn Kategorien (je fünf Kategorien zu den Bereichen *Inhalt* und *Darstellung*), anhand derer die Schreibaufgabe bewertet wurde; zum Anderen die Formalfehler (Fehler der sprachlichen Richtigkeit) Orthographie, Grammatik und Interpunktion. Einzelheiten im ersten Abschnitt.

#### (5) Drei Analyseschritte

1. *Wie weit bleibt beim Übergang von der Erst- zur Zweitkorrektur die Rangordnung der Arbeiten erhalten? Bleiben die beim ersten Mal besseren oder schlechteren Arbeiten auch beim zweiten Mal die besseren oder schlechteren?*

Messgröße: Korrelationskoeffizienten.

2. *Wie groß sind insgesamt die Unterschiede zwischen Erst- und Zweitkorrektur unabhängig davon, ob die Abweichung nach unten oder nach oben erfolgt?*

Messgröße: Absolutbetrag, um den sich Erst- und Zweitkorrektur unterscheiden.

3. *Führt die Zweitkorrektur tendenziell zu besseren oder zu schlechteren Ergebnissen als die Erstkorrektur?*

Messgröße: Die (gerichtete) Differenz "Zweitkorrektur minus Erstkorrektur".

---

## B ERGEBNISSE 1: UNTERSCHIEDE IN DER ANWENDUNG DER ZEHN BEWERTUNGSKATEGORIEN

---

#### (6) Alle Arbeiten: Korrelativer Zusammenhang zwischen Erst- und Zweitkorrektur und absolute Abweichung

**Korrelationen** von Erst- und Zweitkorrektur weisen auf einen nur schwach ausgeprägten Zusammenhang hin; vgl. Tabelle 2.1.1. Eine Betrachtung der **absoluten Abweichungen** zeigt, dass bei den Bewertungsaspekten des *Inhalts* die Erst- von der Zweitkorrektur im Mittel um 2,1 Punkte, bei jenen der *Darstellung* durchschnittlich um 1,3 Punkte und insgesamt um 3,0 Punkte abweicht; vgl. Tabelle 2.1.2.<sup>23</sup>

#### (7) Alle Arbeiten: Die Abweichungen innerhalb von Toleranzbereichen

Abweichungen sind unvermeidlich; tolerieren wir - unter Berücksichtigung der unterschiedlichen Wertebereiche, s.o. - für den Teil *Inhalt* zwei Punkte Abweichung, für den Teil *Darstel-*

---

<sup>23</sup> Bei der Interpretation der Werte sind die unterschiedlichen Wertebereiche zu berücksichtigen: Die möglichen Minima betragen zwar für alle Items und (Teil)Summen 0, die höchstmöglichen jedoch für die Inhaltsitems 2 und somit für die Teilsumme *Inhalt* 10 Punkte, für die Darstellungsitems 1 Punkt, für die Teilsumme *Darstellung* somit 5 und insgesamt für die Schreibaufgabe 15 Punkte.

lung einen Punkt und für die Gesamtsumme drei Punkte, so liegen in diesen drei Toleranzbereichen 66%, 65% und 64%, also jeweils rund zwei Drittel der Vergleichsarbeiten, ein akzeptabler, wenngleich in Zukunft noch auszubauender Anteil; vgl. Abbildung 2.1.3.

### **(8) Alle Arbeiten: Führt die Erst- oder die Zweitkorrektur zu besseren Ergebnissen?**

Für alle Bewertungsaspekte und somit auch für die (Teil)Summen ergeben sich im Mittel Differenzen, die negativ oder Null sind; vgl. Tabelle 2.1.5. Es gibt also mehr Bewertungen, die bei der Erstkorrektur günstiger ausfallen als bei der Zweitkorrektur. Allerdings ist festzuhalten, dass bei einer beträchtlichen Anzahl von Fällen, das Umgekehrte gilt; vgl. Abbildung 2.1.6.

Es gibt eine Tendenz, dass die unterrichtende Lehrkraft etwas besser bewertet als der Zweitkorrektor, aber das ist nicht die Regel.

### **(9) Schulartspezifik: Korrelative Zusammenhänge**

Korrelationen zwischen der Erst- und der Zweitkorrektur weisen darauf hin, dass anscheinend Arbeiten aus der Hauptschule eindeutiger zu bewerten sind als die aus anderen Schularten. Und umgekehrt werden Arbeiten aus dem Gymnasium heterogener bewertet als die der übrigen Schularten; vgl. Tabelle 2.1.7.

Unterscheiden wir zusätzlich, ob der Zweitkorrektor aus derselben oder aus einer anderen Schulart kommt, so ist anzunehmen, dass bei schulartidentischer Korrektur die Ähnlichkeiten (gemessen anhand der Korrelation) größer als bei schulartdifferenter Korrektur sind. Dies aber trifft nicht zu, jedenfalls nicht durchgängig. Die Vorannahme bestätigt sich nur bei den Gesamt- und bei den Hauptschulen, hingegen nicht bei den Realschulen und insbesondere nicht bei den Gymnasien und Berufsschulen; vgl. Tabelle 2.1.8.

Sonderfall Hauptschule: Hier sind die Konstellationen in den Teilbereichen *Inhalt* und *Darstellung* gegensinnig: Die OH-Korrekturen ähneln sich für die inhaltlichen Aspekte mehr, wenn beide Korrektoren aus der Hauptschule kommen, für die darstellerischen Aspekte gilt das Gegenteil, ein Ergebnis, das schulartübergreifend in abgeschwächter Weise ebenfalls gilt.

### **(10) Schulartspezifik: Unterschiede (Absolutbeträge) zwischen Erst- und Zweitkorrektur**

Für die Gesamtbewertung (*Gesamtsumme*) der Schreibaufgabe gilt: Die größte Nähe zwischen Erst- und Zweitkorrektur findet sich für die Arbeiten aus der Hauptschule, die geringste bei jenen aus dem Gymnasium; vgl. Tabelle 2.1.9.

### **(11) Gibt es bei den Kategorien des Bewertungsbereiches *Inhalt* größere Übereinstimmung als bei jenen der *Darstellung*?**

Da das Wertespektrum des Bewertungsbereiches *Inhalt* doppelt so groß ist wie dasjenige der *Darstellung*, ist schwierig zu entscheiden, in welchem der Bereiche die größeren Abweichungen auftreten. Dennoch drängt sich aufgrund des durchgängigen Musters der Eindruck

auf, die Kriterien der *Darstellung* seien divergenter verwendet worden als jene des *Inhalts*; vgl. Tabellen 2.1.2 und 2.1.9. Zwischen den zehn Bewertungskategorien gibt es kaum Unterschiede: Alle sind ähnlich (un)scharf; vgl. ebenfalls Tabelle 2.1.2.

**(12) Schulartspezifik: Führt die Erst- oder die Zweitkorrektur zu besseren Ergebnissen?**

Gehen wir von den Absolutbeträgen über zu den gerichteten Differenzen, so ergibt sich - und dies nahezu durchgängig - das bereits bekannte Bild negativer Differenzen. Mit einer Ausnahme gibt es in allen Schularten die Tendenz, bei der Zweitkorrektur etwas strenger zu bewerten als bei der Erstkorrektur; vgl. Tabelle 2.1.10. Und es gilt zugleich das ebenfalls bereits Festgestellte:

Eine Tendenz ist keine Regel, d.h. die Mittelwerte dürfen nicht vergessen machen, dass es ebenfalls viele Zweitkorrekturen gibt, die zu besseren Ergebnissen als die Erstkorrekturen führen.

Ausnahme Hauptschule: Die Zweitkorrektoren bewerten milder, allerdings nur im Bewertungsbereich *Darstellung*. Im Bereich *Inhalt* hingegen folgt die Zweitkorrektur dem Muster der anderen Schularten, nämlich dem einer tendenziell strengeren Bewertung.

**(13) Schulartspezifik: Macht es einen Unterschied, ob schulartidentisch oder schulartdifferent zweitkorrigiert wird?**

Vermutung: Anzunehmen wäre, dass die schulartidentische Korrektur zu geringeren Abweichungen zwischen der Erst- und der Zweitkorrektur als die schulartdifferente führt; vgl. zum Folgenden die Tabellen 2.1.11 und 2.1.12.

Abweichungen in Absolutbeträgen:

Stimmig mit der Vorannahme sind die Werte für die Gesamt- und die Realschule. Trotz der Unsicherheit, die mit den geringen Fallzahlen bei Haupt- und Berufsschule verbunden ist, ist dennoch bemerkenswert, wie die Konstellationen in den beiden Teilbereichen gegensinnig sind, nämlich hypothesenkonform für *Inhalt*, während für die darstellerischen Aspekte die Abweichungen zwischen Erst- und Zweitkorrektur größer sind bei den schulidentischen Korrekturen. Dieses hypothesenwidrige Resultat gilt (und hier bei hohen Fallzahlen) generell für das Gymnasium.

Unterschiede zwischen Erst- und Zweitkorrektur gemessen in gerichteten Differenzen:

Hier sind die Ergebnisse stark schulartabhängig.

**1. Sind die Bewertungen von Erst- und Zweitkorrektur ähnlicher, wenn der Zweitkorrektor aus derselben Schulart kommt als aus einer anderen?**

	<u>Inhalt</u>	<u>Darstellung</u>	<u>Teil II insgesamt</u>
<b>Insgesamt</b>	( ja ) <sup>24</sup>	---- <sup>25</sup>	----
<b>O</b>	( ja )	( ja )	ja
<b>OH</b>	ja	nein	( ja )
<b>OR</b>	ja	----	ja
<b>OG</b>	nein	nein	nein
<b>OBF</b>	( ja )	nein	----

**2. Gibt es schulartspezifisch größere oder kleinere Differenzen zwischen der Erst- und der Zweitkorrektur?**

Die höchsten Abweichungen finden wir bei den Arbeiten aus den Gymnasien und der Realschule, die geringsten bei jenen aus der Hauptschule.

---

**C ERGEBNISSE 2: DIE FORMALFEHLER**

---

Die Datenbasis für die Auswertung der Formalfehler ist wesentlich schmäler als bei den bisher behandelten Bewertungsaspekten: N=233 Arbeiten mit vollständigen Angaben zu allen Fehlertypen. Die Reichweite der Ergebnisse ist damit eine eingeschränkte. Dies gilt insbesondere für schulartspezifische Analysen, bei denen die Berufsfachschule nicht berücksichtigt werden kann.

**(14) Alle 233 Arbeiten: Korrelative Zusammenhänge**

Die Korrelationskoeffizienten sind zwar deutlich größer als die der Bewertungsaspekte *Inhalt* und *Darstellung*, wesentlich höhere wären aber zu erwarten gewesen, denn abgesehen von immer auftretenden Ungenauigkeiten beim Zählen müssten sich bei den Deutscharbeiten immer dieselben Fehleranzahlen ergeben, da es sich um Fehlertypen handelt, die durch einheitliche und allgemein bekannten Regeln definiert sind; vgl. Tabelle 2.2.2.

Die Rangfolge der Koeffizienten ist plausibel: Der niedrigste Wert ergibt sich für die Grammatikfehler, über die am schwersten Einigkeit zu erzielen ist, während der Fehlerquotient die höchste Korrelation aufweist, also eine aus mehreren Komponenten zusammengesetzte Größe, von denen wir wissen, dass sie tendenziell eine größere Zuverlässigkeit besitzt als ein Einzelmerkmal.

**(15) Alle 233 Arbeiten: Unterschiede zwischen der Erst- und Zweitkorrektur in Absolutbeträgen und (gerichteten) Differenzen**

Die relativ niedrigen Korrelationen finden ihre Bestätigung, gehen wir über zu den absoluten Abweichungen und den Differenzen zwischen den Werten der Erst- und der Zweitkorrektur; vgl. Tabelle 2.2.3.

---

<sup>24</sup> ( ja ): Eingeschränktes Ja.

<sup>25</sup> ----: Kein Unterschied.

Beispiel Umfang der Arbeiten (Anzahl der Wörter): Aus wie viel Wörtern der Text besteht, den Schüler/innen beim Bewältigen von Schreibaufgaben produzieren, wird i.d.R. geschätzt, womit eine gewisse Unsicherheit verbunden ist. Im Mittel beträgt die Abweichung zwischen der Zählung aus der Zweit- und aus der Erstkorrektur 66 Wörter, wenn außen vor bleibt, ob die Abweichung nach unten oder nach oben erfolgt. Berücksichtigen wir diese Richtung, so dass positive und negative Abweichungen gegeneinander verrechnet werden, dann ergibt sich ein Saldo im Schnitt von -53 Wörtern, d.h. tendenziell führt die Zweitkorrektur zu niedrigeren Wortumfängen als die Erstkorrektur. In 36% der Fälle liegen die Zweitschätzungen unter denen der Erstschätzungen; aber bei immerhin rund der Hälfte der Arbeiten kommen die Schätzungen zu denselben Ergebnissen. Dennoch können die Abweichungen zwischen der Erst- und der Zweitkorrektur enorm sein: Beziehen wir die Zweit- auf die Erstangabe, so schwanken diese Anteile von 24% bis 350%!

Bemerkenswert ist, wie selten bei den objektivierbaren *Fehlern der sprachlichen Richtigkeit* Erst- und Zweitkorrektur zum selben Ergebnis kommen.

### **(16) Schulartspezifik: Formalfehler bei Erst- und Zweitkorrektur**

Die schulartspezifische Auswertung weist aufgrund der niedrigen Fallzahlen die Schularten Haupt- und Berufsfachschule nicht gesondert aus.

Insgesamt gilt zweierlei: Es lässt sich kein einheitliches Muster über die Schularten hinweg erkennen - was an den geringen Fallzahlen und der mangelnden Repräsentativität liegen mag -- und die Abweichungen sind für im Prinzip objektive Merkmale erstaunlich hoch und somit sehr unzuverlässige Größen.

### 3.2 SCHLUSSFOLGERUNGEN

#### (a) Bewertungskategorien und Auswertungshinweise

Werden Abweichungen von  $\pm 20\%$  toleriert, d.h. aktuell von  $\pm 3$  Punkten bei insgesamt 15 Punkten, dann liegen rund zwei Drittel aller Arbeiten in diesem Toleranzbereich. Dieser Anteil ist akzeptabel, muss aber vergrößert werden.

Zwar sind die Abweichungen insgesamt nicht dramatisch, wenn die Gesamtbewertung betrachtet wird. Aber bei den einzelnen Bewertungsaspekten treten nahezu durchgängig Unterschiede auf, d.h. häufig kommen Erst- und Zweitkorrektur zu demselben Gesamtergebnis, auch wenn die Einzelaspekte unterschiedlich bewertet wurden. Wünschenswert wäre jedoch eine weitgehende Übereinstimmung auch bei den Teilkriterien.

Hierfür ist es erforderlich - in Anlehnung an Bewertungsmuster aus der Schulforschung - die Bewertungskriterien zu differenzieren und zu präzisieren und dies ausgehend von den Anforderungsmerkmalen der konkreten Darstellungsform. Überschneidungen sind zu vermeiden, Eindeutigkeit ist erforderlich. Wird die Zweiteilung nach *Inhalt* und *Darstellung* beibehalten, dann muss nach den jetzigen Ergebnissen besonderes Augenmerk auf die Aspekte der Gestaltung gelegt werden.

Hier ließen sich ggf. die Aspekte D-3 *Abstrahierende Begriffe* und D-4 *Wortwahl* zu einer Kategorie *Sachgerechter Ausdruck* zusammenfassen. Dies erscheint auch deswegen angezeigt, weil der Abstraktionsgrad der Ausführungen sich eher auf der inhaltlichen Ebene erkennen lässt - dort jedoch fehlt dieses Kriterium und könnte z.B. I-4 *Gewichtung der Argumente* ersetzen, das faktisch von geringer Bedeutung ist, denn diesbezügliche Formulierungen der Schüler/innen erschöpfen sich in Wendungen wie "Der wichtigste Punkt aber ist ...".

Von den inhaltlichen Kriterien scheint I-1 *These in eigenen Worten* am uneinheitlichsten verstanden worden zu sein. Für die Einleitung einer Problemerkörterung, und um eine solche ging es bei der Schreibaufgabe, werden die Schüler/innen angehalten, die These zu erläutern und klar auf den Punkt zu bringen, was nicht dasselbe ist wie die These mit eigenen Worten zu formulieren. Im Einklang mit dieser unterrichtlichen Vorgabe könnte I-1 zu *Adäquate Erläuterung der These* verändert werden.

Über die skizzierten Modifikationen hinaus sind zwei weitere Ansatzpunkte, zu einer weitgehend einheitlichen und vergleichbaren Bewertung zu kommen, denkbar.

1. Die korrigierenden Lehrkräfte werden gebeten, zunächst eine vorläufige Gesamtpunktzahl, hier zwischen 0 und 15 Punkten, zu bestimmen, die ihrem Eindruck von der Güte der zu bewertenden Arbeit am besten entspricht. Danach sollen sie die Punkte auf die einzelnen Kategorien, hier I-1 ff. und D-1 ff., nach Maßgabe der jeweils maximal möglichen Punkte verteilen.

Die systematische Durchsicht der Aspekte verändert ggf. den Gesamteindruck und die Gesamtpunktzahl, was wiederum eine neuerliche Aufteilung nach sich zieht. Ein derartiger schleifenartiger Bewertungsmodus dürfte zumindest die Reliabilität und in gewissem Maße auch die Validität erhöhen.

2. Ein qualitativ anderer Ansatz könnte der Konstruktion der Schreibaufgabe selber gelten, also gewissermaßen nicht auf die korrigierenden Lehrkräfte, sondern auf die schreibenden Schüler/innen zielen, indem die Schreibaufgabe durch Teilvorgaben konkretisiert und präzisiert wird. Für jede Teilaufgabe ließe sich hierdurch relativ genau beurteilen, ob und wie weit ein/e Schüler/in den Anforderungen genügt hat. Damit allerdings näherten sich von Struktur und Format die beiden Teile der Vergleichsarbeit einander an<sup>26</sup> und das Fähigkeitspektrum, das in den Vergleichsarbeiten zum Tragen kommt, wird eingeeengt.

### **(b) Korrektur durch die unterrichtende Lehrkraft**

Werden Aufwand und Ertrag ins Verhältnis gesetzt, so ist es zum jetzigen Zeitpunkt gerechtfertigt, die bisherige Praxis Vergleichs- wie Klassenarbeiten von der unterrichtenden Lehrkraft korrigieren zu lassen, beizubehalten. Zwar fällt im Mittel die Erstkorrektur etwas milder als die Zweitkorrektur aus, aber in zahlreichen Fällen gilt das Gegenteil, so dass sich keine generelle Regel nachsichtiger Erstkorrektur aufstellen lässt. Im Übrigen lässt sich nicht entscheiden, ob die Erst- oder die Zweitkorrektur objektiv richtiger ist. Der Gewinn einer mit erheblichem Aufwand verbundenen Umstellung wäre somit zweifelhaft.

Zugleich lässt sich daraus ableiten, dass die bislang gewonnenen landesweiten Bezugswerte aussagekräftig und verlässlich sind. Die Abweichungen zwischen Erst- und Zweitkorrektur gleichen sich zum größten Teil aus.

### **(c) Schulartspezifische Anstrengungen**

Auffällig sind die Besonderheiten, die vor allem das Gymnasium, aber auch die Realschule auszeichnen. Arbeiten aus dem Gymnasium werden heterogener bewertet als die Arbeiten aus anderen Schularten - und dies gilt verstärkt dann, wenn der Zweitkorrektor ebenfalls aus dem Gymnasium kommt. In etwas abgeschwächter Form bleibt diese Aussage ebenfalls für die Realschule gültig.

Ziel muss demnach sein, vor allem in diesen beiden Schularten ein gemeinsames Verständnis darüber herzustellen, welche Kriterien für die Bewertung von Schreibaufgaben gelten und wie diese anzuwenden seien.

### **(d) Die Formalfehler**

Die Untersuchung hat gezeigt, dass diese an sich objektivierbaren Fehlertypen zu äußerst divergenten Resultaten führen. Die Bewertung anhand dieser Kriterien ist sehr unzuverlässig und kann demnach ohne Einbußen abgeschafft werden und dies umso mehr, als es kaum einen Zusammenhang zwischen den Formalfehlern und den inhaltlichen Bewertungskriterien gibt, wie der Anhang A4 zeigt.

---

<sup>26</sup> Teil I Lesen - Mit Texten und Medien umgehen besteht aus Multiple-Choice-Aufgaben und Fragen, die in kurzen Sätzen oder Stichwörtern zu beantworten sind. Die zu vergebenden Punktzahlen reichten i.d.R. bis maximal 2 Punkten.

**(e) Weitere Zweitkorrekturen**

In den nächsten Jahren müssen weiterhin stichprobenartig Zweitkorrekturen der Schreibaufgaben vorgenommen werden, um überprüfen zu können, ob die Entwicklung in die gewünschte Richtung verläuft. (Dabei sollte der jetzt realisierte Ansatz, die Arbeiten über alle Schularten zur Zweitkorrektur zu verteilen, beibehalten werden, um als willkommenen Nebeneffekt zumindest den davon betroffenen Lehrkräften neben den landesweiten Bezugswerten auch einen Vergleich mit konkreten Arbeiten, die nicht aus ihrer Klasse stammen, zu ermöglichen.) Zugleich ist eine Vergrößerung der Fallzahlen anzustreben, um die im Anhang A3 skizzierten weiteren Auswertungsschritte unternehmen zu können, die ihrerseits daraus folgende Maßnahmen punktgenauer zu gestalten helfen könnten.

## ANHANG

**A1 Tabelle:** Die Teilpunktsummen zu *Inhalt* und *Darstellung* sowie die Gesamtpunktzahl für den Teil II. Streuungen der Unterschiede zwischen den Bewertungen aus der Erst- und der Zweitkorrektur differenziert nach Schulart, aus denen die Arbeiten stammen, und nach Schulart der Zweitkorrektur. (Anggegeben sind die Streuungen der Absolutbeträge und - in Klammern - der (gerichteten) Differenzen.)

Arbeiten	N	<i>Inhalt</i>	<i>Darstellung</i>	<i>Gesamtsumme</i>
<b>Alle Arbeiten</b>	746	1,7 (2,6)	1,1 (1,7)	2,4 (3,7)
<b>K1=S=K2</b>	306	1,7 (2,5)	1,2 (1,8)	2,5 (3,8)
<b>K1=S/K2≠S</b>	440	1,7 (2,6)	1,0 (1,6)	2,4 (3,7)
aus der <b>Gesamtschule</b>	202	1,6 (2,4)	1,0 (1,6)	2,2 (3,5)
<b>K1=O=K2</b>	71	1,6 (2,3)	0,9 (1,5)	2,2 (3,3)
<b>K1=O/K2≠O</b>	131	1,6 (2,5)	1,1 (1,7)	2,2 (3,7)
aus der <b>Hauptschule</b>	54	1,7 (2,4)	1,0 (1,3)	2,1 (3,1)
<b>K1=OH=K2</b>	30	1,4 (2,0)	1,0 (1,6)	2,3 (3,1)
<b>K1=OH/K2≠OH</b>	24	2,0 (2,9)	0,8 (0,9)	2,0 (3,1)
aus der <b>Realschule</b>	172	1,7 (2,7)	1,1 (1,6)	2,4 (3,8)
<b>K1=OR=K2</b>	75	1,6 (2,5)	1,2 (1,7)	2,3 (3,7)
<b>K1=OR/K2≠OR</b>	97	1,8 (2,8)	1,1 (1,6)	2,5 (3,9)
aus dem <b>Gymnasium</b>	252	1,7 (2,7)	1,2 (1,6)	2,6 (3,9)
<b>K1=OG=K2</b>	92	1,9 (2,9)	1,4 (1,8)	2,9 (4,1)
<b>K1=OG/K2≠OG</b>	160	1,6 (2,6)	1,0 (1,5)	2,4 (3,7)
aus der <b>Berufsfachschule</b>	66	1,4 (2,2)	1,3 (2,1)	2,4 (3,7)
<b>K1=OBF=K2</b>	38	1,3 (2,1)	1,4 (2,3)	2,5 (3,8)
<b>K1=OBF/K2≠OBF</b>	28	1,4 (2,4)	1,0 (1,7)	2,3 (3,7)

**A2 Tabelle:** Die Teilpunktsummen zu *Inhalt* und zur *Darstellung* sowie die Gesamtpunktzahl für den Teil II. Mittelwerte und Streuungen der Bewertungen aus der Erst- und der Zweitkorrektur differenziert nach Schularten.  
(Streuungen in Klammern. Oben stehen jeweils die Werte der Erst-, darunter jene der Zweitkorrektur.)

Arbeiten	N	<i>Inhalt</i>	<i>Darstellung</i>	<i>Gesamtsumme</i>
<b>Alle Arbeiten</b>	746	5,5 (2,8) 4,7 (2,4)	2,9 (1,6) 2,7 (1,6)	8,4 (4,0) 7,4 (3,7)
<b>K1=S=K2</b>	306	5,1 (2,7) 4,5 (2,4)	2,9 (1,6) 2,6 (1,6)	8,0 (4,0) 7,0 (3,6)
<b>K1=S/K2≠S</b>	440	5,7 (2,7) 4,9 (2,5)	3,0 (1,6) 2,8 (1,6)	8,7 (4,0) 7,7 (3,8)
aus der <b>Gesamtschule</b>	202	5,0 (2,7) 4,5 (2,5)	2,5 (1,5) 2,5 (1,5)	7,5 (3,9) 7,0 (3,7)
<b>K1=O=K2</b>	71	4,9 (2,7) 4,3 (2,5)	2,6 (1,6) 2,4 (1,3)	7,5 (3,9) 6,7 (3,4)
<b>K1=O/K2≠O</b>	131	5,0 (2,7) 4,6 (2,4)	2,5 (1,5) 2,6 (1,6)	7,5 (3,9) 7,2 (3,8)
aus der <b>Hauptschule</b>	54	3,2 (2,9) 2,9 (2,3)	1,4 (1,3) 1,9 (1,6)	4,6 (3,7) 4,8 (3,6)
<b>K1=OH=K2</b>	30	3,0 (2,8) 3,0 (2,5)	1,6 (1,4) 2,0 (1,8)	4,6 (4,0) 5,0 (4,0)
<b>K1=OH/K2≠OH</b>	24	3,4 (2,6) 2,7 (2,1)	1,2 (1,2) 1,8 (1,3)	4,6 (3,5) 4,5 (3,0)
aus der <b>Realschule</b>	172	5,4 (2,7) 4,7 (2,4)	2,9 (1,5) 2,5 (1,5)	8,3 (3,8) 7,2 (3,6)
<b>K1=OR=K2</b>	75	5,2 (2,6) 4,8 (2,1)	2,8 (1,5) 2,6 (1,5)	7,9 (3,7) 7,4 (3,3)
<b>K1=OR/K2≠OR</b>	97	5,7 (2,7) 4,6 (2,5)	2,9 (1,4) 2,5 (1,6)	8,6 (3,9) 7,1 (3,7)
aus dem <b>Gymnasium</b>	252	6,9 (2,1) 5,7 (2,2)	3,8 (1,2) 3,2 (1,5)	10,7 (3,1) 8,9 (3,4)
<b>K1=OG=K2</b>	92	6,8 (2,0) 5,4 (2,1)	3,9 (1,2) 3,0 (1,6)	10,7 (2,9) 8,4 (3,3)
<b>K1=OG/K2≠OG</b>	160	6,9 (2,2) 5,9 (2,2)	3,8 (1,3) 3,4 (1,4)	10,7 (3,2) 9,3 (3,4)
aus der <b>Berufsfachschule</b>	66	3,6 (2,4) 3,5 (1,9)	2,3 (1,7) 2,2 (1,5)	5,9 (3,6) 5,6 (3,1)
<b>K1=OBF=K2</b>	38	3,2 (2,1) 3,1 (1,7)	2,3 (1,8) 2,1 (1,5)	5,5 (3,2) 5,2 (3,0)
<b>K1=OBF/K2≠OBF</b>	28	4,1 (2,6) 3,9 (2,1)	2,3 (1,7) 2,3 (1,4)	6,4 (4,0) 6,2 (3,2)

### A3 PERSPEKTIVISCHES

Welche Möglichkeiten einer Feinanalyse sich ergäben, wären die Fallzahlen größer, kann nur an zwei Beispielen demonstriert werden, bei denen die Kategorie schulartdifferenter Korrekturen - hier für Arbeiten aus der Gesamt- und der Realschule ( $K1=O/K2\neq O$  und  $K1=OR/K2\neq OR$ ) - ausdifferenziert wird - im nachstehenden Beispiel für Zweitkorrektoren aus dem Gymnasium, also  $K1=O/K2=OG$  und  $K1=OR/K2=OG$ . Tabelle A3.1 liefert die Ergebnisse der entsprechenden Berechnungen.

**A3.1 Tabelle:** Die Teilpunktsummen zu *Inhalt* und zur *Darstellung* sowie die Gesamtpunktzahl für den Teil II. Unterschiede in den Bewertungen von Erst- und Zweitkorrektur für Arbeiten aus der Gesamt- und der Realschule bei Zweitkorrektoren aus dem Gymnasium. (Angegeben werden jeweils die Mittelwerte der Unterschiede zwischen Erst- und Zweitkorrektur in Absolutbeträgen und - jeweils darunter - den gerichteten Differenzen. In Klammern stehen zum Vergleich die Werte  $K1-O\neq K2-O$  und  $K1-OR\neq K2-OR$  aus den Tabellen 2.1.11 und 2.1.12.)

Arbeiten	N	<i>Inhalt</i>	<i>Darstellung</i>	<i>Gesamtsumme</i>
Alle Arbeiten $K1=S/K2\neq S$	746	( 2,2)	( 1,2)	( 3,0)
		(-0,8)	(-0,2)	(-0,9)
Gesamtschule $K1=O/K2=OG$	46	2,1 ( 2,0)	1,2 ( 1,3)	3,1 ( 2,9)
		-0,4 (-0,3)	-0,2 ( 0,1)	-0,7 (-0,2)
Realschule $K1=OR/K2=OG$	46	2,8 ( 2,4)	1,4 ( 1,3)	3,7 ( 3,3)
		-1,5 (-1,0)	-0,5 (-0,4)	-2,0 (-1,5)

Es zeigt sich, dass bei den Realschularbeiten die Bewertungen der gymnasialen Zweitkorrektoren bei den Inhaltsaspekten im Mittel Abweichungen von der Erstkorrektur aufweisen, die über dem Durchschnitt aller Zweitkorrekturen liegen<sup>27</sup>, Abweichungen, die sich in der Gesamtsumme niederschlägt. Dies ist weder bei den Darstellungskriterien noch bei den Gesamtschularbeiten der Fall. Dieses Beispiel macht deutlich, dass unterschiedliche Konstellationen in den Teilen zu den globalen Ergebnissen führen, die Gegenstand der Darstellung waren. Ziel einer Nachfolgeuntersuchung müsste es demnach sein, die Fallzahlen derart zu vergrößern, dass in der Tabelle 1.2 jede Zelle hinreichend besetzt um, um zuverlässige Detailanalysen durchführen zu können.

<sup>27</sup> Wobei diese Abweichungen zwischen den gymnasialen und den anderen Zweitkorrekturen noch größer sind, als in der Tabelle 2.1.13 angegeben, denn die gymnasialen Bewertungen sind in den Klammerwerten, die ja aus den Tabellen 2.1.11 und 2.1.12. stammen, mit enthalten. Da es an dieser Stelle nur darum ging, dass prinzipielle Analysepotenzial des Untersuchungsdesigns zu verdeutlichen, wurde der Mehraufwand getrennter Berechnung gemieden.

#### A4 ZUM ZUSAMMENHANG ZWISCHEN DEN FORMALFEHLERN UND DEN BEWERTUNGEN NACH DEN ASPEKTEN *INHALT* UND *DARSTELLUNG*

Der (statistische) Zusammenhang zwischen den Formalfehlern und den Bewertungen anhand der Aspekte zum *Inhalt* und zur *Darstellung* ist sowohl für die Erst- wie für die Zweitkorrektur gering, wie die Tabelle A4.1 zeigt.

**A4.1 Tabelle:** Die Formalfehler (Fehler der sprachlichen Richtigkeit) und ihr (statistischer) Zusammenhang mit den Bewertungsaspekten *Inhalt* und *Darstellung*. Korrelationen<sup>28</sup> für die Erst- und die Zweitkorrektur. (Grundlage N=233 Arbeiten; K1 bzw. K2: Werte für die Erst- bzw. Zweitkorrektur.)

Punktzahl →			<i>Inhalt</i>	<i>Darstellung</i>	Teil II
F-1: Wortanzahl	K1:		.24	.16	.24
	K2:		.30	.15	.26
F-5: Gesamtfehlerzahl	K1:		-.23	-.41	-.32
	K2:		-.30	-.37	-.36
F-6: Fehlerquotient	K1:		-.41	-.50	-.49
	K2:		-.43	-.42	-.47

Bei beiden Korrekturen gibt es einen sehr schwachen positiven Zusammenhang zwischen dem Umfang einer Arbeit und deren positiver Bewertung (tendenziell: je länger die Arbeit, desto besser deren Bewertung) - bemerkenswerterweise stärker ausgeprägt für den Bereich *Inhalt* als für den Bereich *Darstellung*; das Gegenteil zu erwarten hätte nahe gelegen.

Die Zusammenhänge zwischen den Punktbewertungen und den Fehlern ist stärker und sachgemäß negativ: Tendenziell gilt, dass je mehr Fehler gemacht werden, desto negativer eine Arbeit bewertet wird. Die Korrelationen sind allerdings auch hier nicht sehr hoch.

<sup>28</sup> SPEARMANs  $\rho$