Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur
http://www.didaktik-deutsch.de
14. Jahrgang 2009 – ISSN 1431-4355
Schneider Verlag Hohengehren GmbH

Peter Birkel

RECHTSCHREIBLEISTUNG IM DIKTAT – EINE OBJEKTIV BEURTEILBARE LEISTUNG?

In: Didaktik Deutsch. Jg. 14. H. 27. S. 5-32.

Peter Birkel

RECHTSCHREIBLEISTUNG IM DIKTAT – EINE OBJEKTIV BEURTEILBARE LEISTUNG?

1. Einleitung

Diktate schreiben zu lassen gehört einerseits mit zu den bestens bekannten Tätigkeiten eines Deutschlehrers in der Grundschule und der Sekundarstufe I, obwohl sie in fast allen Bundesländern aus den Vorgaben der Lehrpläne und Richtlinien verschwunden sind (Brinkmann 2004, S. 11). Neben dem traditionellen Diktat, bei dem die Schüler einen zusammenhängenden Text nach Gehör zu schreiben haben, etablierten sich verschiedene alternative Diktatformen, wie sie z.B. bei Korn (2005, Kap. 6) beschrieben sind mit dem Impetus, weniger Kontroll- und Selektionsinstrument zu sein als eher der Lernförderung zu dienen. Bei allen Diktatformen wird davon ausgegangen, dass eigentlich jeder Lehrer, gleichgültig ob er das Fach Deutsch studiert hat oder nicht, in der Lage sei, die Korrektur des Diktats ordnungsgemäß ausführen zu können. Zumindest unter Inanspruchnahme eines Dudens müsse doch eindeutig zu entscheiden sein, ob ein Wort richtig oder falsch geschrieben wurde. Folglich müsse auch das individuelle Ergebnis eines Schülers anhand der Fehlerzahl recht objektiv feststellbar sein. Verschiedene Lehrer müssten zum praktisch gleichen Ergebnis kommen.

Andererseits hat die Deutschdidaktik, beginnend bereits in den 70er Jahren, das Diktatschreiben zunehmend kritisch diskutiert. Die Diktatkritik reicht dabei bis ins 19. Jahrhundert zurück (s. Fix 1994, 2004). Die Kritik bezieht sich vor allem darauf, dass Diktate als Lernkontrollen bei Weitem nicht so effektiv wirken wie angenommen, der Lerneffekt dabei gering bis kaum vorhanden und mit anderen Mitteln viel besser erreichbar sei, dass sie in einer psychisch belasteten Ausnahmesituation zu schreiben sind und die Entwicklung von angemessenen Strategien für den Gebrauch der Schriftsprache anders viel eher zu leisten sei (Spitta 1976, Fix 1994, Menzel 1997). Die Ablehnung des Diktats durch die Fachdidaktiker scheint aber weder die älteren Lehrkräfte in den Schulen davon abzuhalten, weiterhin mit Diktaten die Rechtschreibfähigkeit ihrer Schüler überprüfen zu wollen, noch die jüngeren Hochschulabsolventen, sich dem Diktat des Diktats zu entziehen. Fix (1994) benennt einige Gründe, warum es die neuen Einsichten zum Diktatschreiben so schwer haben, sich in der Praxis durchzusetzen. Einer der wichtigsten ist sicher im arbeitsökonomischen Vorteil zu sehen, aber auch die konservative Haltung von Gesellschaft und Lehrerkollegien, die Selektionsfunktion der Schule und die disziplinarische Funktion des Diktats können eine Rolle spielen. Lehrer scheinen sehr resistent gegen die Argumente zu sein, Diktate doch eher abzuschaffen (vgl. Adrion 1984). Trotz aller Bemühungen, die Vormachtstellung des traditionellen Diktats im Rechtschreibunterricht zu beenden, konstatiert Leßmann (2004, 33): „Das Diktat ist noch immer mächtigster Faktor bei der Notenfindung im Bereich Rechtschreiben – und raubt damit anderen, weitaus bedeutsameren Teilleistungen der Rechtschreibung den ihnen gebührenden Stellenwert in Bewertung und Unterricht.“

Das Dilemma von Lehrkräften, die um die Funktion von alternativen Möglichkeiten der Rechtschreibförderung wissen, trotzdem aber Diktate schreiben müssen, weil z.B. ein Beschluss der Gesamtlehrerkonferenz vorliegt, beschreibt Hiller (2004, 38) anschaulich. Sie versucht die Diktatarbeit wenigstens didaktisch so aufzuwerten, dass die Schüler eine gute Chance haben, aus ihrem Unterricht eine Menge zu lernen. Ihre Grundeinstellung beschreibt sie mit dem Satz: „Ich ließ mir meine Überzeugung nicht nehmen, dass selbst der Umgang mit einem Diktat dennoch Lernchancen eröffnen kann."

Viele Lehrerkollegien gehen so selbstverständlich davon aus, dass Diktate zu korrigieren und zu beurteilen zu „objektiven", d.h. subjektive Einflüsse ausschließenden Ergebnissen führt. Die Äußerungen der Lehramtsstudenten unserer Hochschule lassen ebenfalls auf eine ähnliche Einstellung schließen. Trotzdem gibt es fachdidaktische Publikationen, die genau auch die Objektivität von reinen Diktaten bezweifeln (z.B. Korn 2005), auch wenn sie diese nicht direkt überprüft haben. Offenbar kam noch niemand auf die Idee, die Gültigkeit dieser Annahme zu überprüfen. Genau hier setzt diese Untersuchung an.

Wenn man empirische Literatur zur Leistungsbeurteilung in Diktaten sucht, stellt man fest, dass es dazu fast keine Untersuchungen gibt. Man stößt dabei allerdings auf eine sehr alte, methodisch den heutigen Ansprüchen nicht gerecht werdende Publikation aus dem Jahr 1928. Damals untersuchte Maria Zillig, bei welchen Schülern Lehrer Diktatfehler eventuell übersehen würden. Sie fand heraus, dass die Lehrer vor allem bei guten Schülern einmal Fehler übersehen, während sie bei schlechten Schülern tatsächlich praktisch alle Fehler identifizierten. Möglicherweise führte die Erwartung, das Diktat eines rechtschreibsicheren Schülers korrigieren zu müssen, dazu, das Diktat schneller überfliegen zu können. Das aber erhöht die Wahrscheinlichkeit, auch einmal einen Fehler zu übersehen. Bei den rechtschreibschwachen Schülern dagegen wird unterstellt, dass praktisch jedes Wort einen Fehler enthalten könnte. Hier muss der Lehrer aufmerksam und langsam lesen, um keinen Fehler zu übersehen. So wird hier praktisch jeder Fehler identifiziert.

Die Untersuchung von Zillig wird seither immer wieder zitiert, obwohl insgesamt der Effekt der Voreinstellung der Lehrer eher geringfügig ausfiel. Eine zufallskritische Beurteilung der Unterschiede wurde nicht vorgenommen. Eine Replikationsstudie zu diesem Phänomen steht bisher aus. So kann diese frühe Untersuchung von allen Seiten so interpretiert werden, wie es den eigenen Intentionen dient. Zumindest der Verdacht, dass auch die Korrektur von Diktaten Einflüssen unterliegen kann, die nichts mit der Rechtschreibleistung selbst zu tun haben, ist aber seither auch nicht ausgeräumt worden.

Thiel & Valtin (2002) interessierten sich für die Fragen, ob die Noten verschiedener Fächer vergleichbar seien, die Strenge der Benotungen im Laufe der Schulzeit zunähme, Jungen und Mädchen vergleichbar benotet würden und Noten aus verschiedenen Klassen vergleichbar seien.

Im Hinblick auf die Strenge der Noten in verschiedenen Fächern bestätigen sie im Wesentlichen die Ergebnisse, die Ingenkamp (1971) bereits berichtet hat. Die selektiven Fächer (Deutsch, Mathe, Fremdsprache) werden strenger benotet als die

nicht selektiven Fächer im musischen und technischen Bereich. Dabei wird ab Klasse 3 das Rechtschreiben am strengsten benotet, gefolgt von Mathematik. Die Notendifferenz zu Fächern wie Kunst oder Sport wächst im Laufe der Jahre von einer halben Notenstufe in Klasse 3 auf mehr als eine ganze Notenstufe in Klasse 6 an. Das bedeutet zudem, dass die Benotungen nicht in allen Fächern gleich strenger werden im Laufe der schulischen Karriere. Das trifft in erster Linie zu auf die Rechtschreibnoten und die Fächer Deutsch und Mathematik. Bekommen in Klasse 2 noch fast 70% der Schüler die Noten 1 und 2 im Rechtschreiben, sind es in Klasse 6 nur noch etwa 35%, und die Verteilung erstreckt sich dann fast glockenförmig über die gesamte Notenskala.

Bezüglich der speziellen Fähigkeiten von Jungen und Mädchen zeigt sich, dass die Mädchen vor allem in den sprachlichen Bereichen und den Fächern, bei denen es auf möglichst angepasstes Verhalten ankommt (Kunst, Musik), bessere Noten bekommen. Spezielle Jungendomäne ist die Mathematik, auch wenn das bei Thiel und Valtin (2002) nur bei den Drittklässlern signifikant wird. Da zu diesen Fächern auch objektive Leistungstestergebnisse vorlagen, ließ sich feststellen, dass die meist bessere Bewertung der Mädchen mit besseren Testergebnissen korrelierte. Trotzdem scheinen die Ergebnisse von Regressionsrechnungen darauf hinzudeuten, dass die Mädchen beim Lesen und Rechtschreiben in einigen Fällen bessere Zensuren als zu erwarten erhielten.

Bei der Vergleichbarkeit der Noten in verschiedenen Klassen wurde ebenfalls im Wesentlichen das Ergebnis von Ingenkamp (1971) repliziert. Das unterschiedliche Leistungsniveau der verschiedenen Klassen führte dazu, dass im Extremfall die schlechtesten Schüler einer leistungsstarken Klasse besser waren als alle Schüler in der leistungsschwächsten Klasse, auch wenn in beiden Klassen die Zensurenverteilungen praktisch gleich ausfielen. Was hier am Bespiel der Mathematik aufgezeigt wird, gilt sicher auch für die Beurteilung der Rechtschreibfähigkeit. Noten können stark abhängen von der eher zufälligen Zugehörigkeit zu einer bestimmten Klasse. Was noch recht gut klappt, ist die Abbildung der relativen Fähigkeit eines Schülers innerhalb einer Klasse auf der Skala des objektiven Tests, auch wenn es in Einzelfällen hier zu Verzerrungen kommen kann. Auch das wurde an einem Mathematiktest verdeutlicht, aber die Gültigkeit dieses Ergebnisses wird im Grunde auch für die Benotung der Rechtschreibleistungen unterstellt.

Im Rahmen einer Lehrveranstaltung an der PH Weingarten wollten sich die Studierenden einmal der Fragestellung annehmen, wie objektiv die Auswertung und Benotung eines Diktats vorgenommen werden kann. Zwar war die allgemeine Erwartung eher so, dass man die hohe Objektivität der Diktatzensuren bestätigt finden würde, aber es gab immerhin auch einige Zweifler, die es für möglich hielten, dass diese Zensuren vielleicht doch nicht so objektiv seien, wie immer beschworen. Hier verhielten sich die Studierenden, wie es auch der Erfahrung von Menzel entspricht, der schreibt (1997, 16f): „Für viele ist das Diktat das zweckmäßigste Instrument zur Überprüfung der Rechtschreibfähigkeit; für manche ist es sogar das einzige. Seine Beliebtheit dankt es wohl der Einfachheit seiner Durchführung und der (scheinbaren) Eindeutigkeit bei der Fehlerfeststellung und Fehlerzählung. Darüber hinaus gilt

es bei vielen als Instrument, das zu ‚objektiven', zumindest vergleichbaren Ergebnissen führt."

2. Planung und Durchführung der Untersuchung

Die Objektivität der Diktatzensuren sollte sich in einer guten Beurteilerübereinstimmung zeigen, wenn viele Lehrkräfte die Diktate korrigieren und benoten. Man brauchte also zunächst einmal ein Diktat, das man zu diesem Zweck den Lehrkräften zur Korrektur übergeben könnte. Das wurde gefunden, als eine Studierendengruppe im Wochentagspraktikum[1] an einer Ausbildungsschule von einem ausführlich vorbereiteten Diktat einer siebten Hauptschulklasse erfuhr. Der Klassenlehrer hatte das Andenken an die Schlacht von Jena und Auerstedt im Jahr 1806 zum Anlass genommen, in einem fächerübergreifenden Unterricht auf die historischen Ereignisse einzugehen. Grundlage war ein Zeitungsartikel, der auf die Überheblichkeit des preußischen Königreichs verwies, die mit dazu beitrug, dass die napoleonischen Truppen diese Schlacht gewannen und die Flucht des preußischen Königs nach Memel erzwangen (Anhang 3 auf der Internetseite von „Didaktik Deutsch"). Im Deutschunterricht wurde der Artikel mehrfach gelesen und auf für die Kinder fremde Begriffe hin untersucht. Klärung der Begriffe, intensive Arbeit an der Kleinschreibung von Adjektiven in Verbindung mit Substantiven (kam in dem Artikel gehäuft vor) und mehrfaches Schreiben sollten zu einer gewissen Sicherheit in der richtigen Schreibung führen. Schließlich wurde den Schülern an einem Freitag mitgeteilt, dass aus diesem relativ langen Zeitungsartikel ein Diktattext erarbeitet würde, der am darauf folgenden Montag zu schreiben sei. Nach der Rückgabe des Diktats wurde das Thema „Kleinschreibung von Adjektiven" weiter vertieft.

Aus den korrigierten Diktaten wurden dann drei ausgewählt. Sie waren vom Deutschlehrer mit den Noten „gut" (Daniela)[2], „ausreichend" (Paul) und „ungenügend" (Markus) beurteilt worden. Die drei Diktate wurden in der Originalhandschrift kopiert, nachdem alle Korrekturzeichen gelöscht waren. Jeweils zwei wurden den Lehrern dann zur Korrektur und Beurteilung der Rechtschreibleistung in der Weise vorgelegt, dass etwa die Hälfte der Lehrer die Diktate von Markus und Paul (Version 1), die andere Hälfte die von Daniela und Paul (Version 2) erhielt.

Um den beurteilenden Lehrern einen Einblick in die unterrichtliche Einbettung des Diktats zu vermitteln, wurden ihnen der Zeitungsartikel in Kopie und eine kurze Beschreibung der unterrichtlichen Aktivitäten (s. Anhang 2 im Internet) zugeleitet. Zudem wurde erfragt:

- Geschlecht der Lehrkraft
- Anzahl der Dienstjahre

[1] An der PH Weingarten müssen Wochentagspraktika an Ausbildungsschulen in der Form absolviert werden, dass die Studierenden zusammen mit Ausbildungslehrern und den Lehrenden jeweils am Mittwoch vormittags Unterricht für die Schüler vorbereiten, durchführen und danach reflektieren.

[2] Die Schülernamen wurden so geändert, dass das Geschlecht der Schüler erkennbar bleibt. Die Originaldiktate sind in Anhang 5 – 7 einsehbar.

- Unterrrichtserfahrung im Fach Deutsch
- Lehrtätigkeit überwiegend an Grund- oder Hauptschule
- Schwierigkeitseinschätzung des Diktattextes

Schließlich sollten die Lehrkräfte auf einem vorbereiteten Bogen angeben:

- Gesamtfehlerzahl
- Note, die sie erteilen würden.

Weiters wurde Platz angeboten, um Kommentare zur eigenen Bewertung oder für den Schüler abgeben zu können, wovon aber nur in Ausnahmefällen Gebrauch gemacht wurde.

Mit diesen Materialien schwärmten die Studierenden aus[3], um möglichst heimatortnah Lehrer an Grund- und Hauptschulen zu finden, die bereit waren, jeweils zwei Diktate zu bearbeiten. Damit war ein relativ großer Einzugsbereich der beteiligten Lehrkräfte im Land Baden-Württemberg gesichert. An einer Schule sollte immer nur eine Version der Diktatzusammenstellung eingesetzt werden.

3. Fragestellungen und Hypothesen

Aus der Anlage der Untersuchung ist unschwer zu erkennen, dass folgende Fragestellungen untersucht werden sollten:

1. *Wie groß ist die Übereinstimmung der Lehrerschaft bei der Anzahl der identifizierten Fehler und bei den daraufhin erteilten Noten?* Die einhellige Erwartung der Seminarteilnehmer war, dass es doch eindeutig anhand eines Dudens entscheidbar sei, ob ein Wort richtig oder falsch geschrieben sei. Darum favorisierte man die Hypothesen: ***1a) Von geringen Schwankungen abgesehen werden die Korrekturergebnisse in etwa übereinstimmen. 1b) Auch die Zensuren werden sich kaum unterscheiden, wenngleich hier über den subjektiven Ermessensspielraum der Lehrkraft etwas mehr Streuung möglich ist.*** Dass diese Annahme etwas naiv ist, weil schon in der Vergangenheit darauf hingewiesen wurde, wie sehr sich Lehrer bei der Auswertung von Diktaten in der Zählung der Fehler unterscheiden können (Hinweise bei Menzel 1997), wollten die Studierenden nicht gelten lassen.

2. *Wie stark weichen die Beurteilungen ab von denen, die der Deutschlehrer im Original erteilt hatte?* Die bisherigen Untersuchungen zur Zensurengebung an der PH Weingarten (Birkel & Birkel 2002, Birkel 2003, Birkel 2005) zeigten übereinstimmend, dass der Mittelwert der abgegebenen Beurteilungen ziemlich genau der im Original erteilten Zensur entsprach. Auf diesem Hintergrund war unsere Hypothese: ***2) Die Mittelwerte der Beurteilungen weichen nicht wesentlich von den Originalbewertungen ab.***

3. *Wird das Diktat von Paul unabhängig von dem Kontextdiktat (Daniela oder Markus) etwa gleich beurteilt?* Da bei der Korrektur der Diktate eine eher große Übereinstimmung erwartet wurde, neigten die Studierenden dazu, hier keine Unterschiede zu erwarten. Hypothese ***3a): Bei der Fehlerfindung zeigt sich kein Kon-***

[3] Für die Teilaufbereitung der Daten von mindestens fünf Lehrkräften erhielten die Studierenden einen Seminarschein (Modul-2-Schein).

texteffekt. Auch bei der Bewertung neigten die Studierenden dazu, höchstens zufällige Unterschiede zu erwarten. Andererseits legte der nachgewiesene Kontrasteffekt bei mündlichen Prüfungen (Birkel 1978, 1984) nahe, zumindest bei der Bewertung von Pauls Diktat einen Kontexteffekt zu erwarten. Man einigte sich auf die Hypothese: ***3b) Pauls Diktat wird im Kontext mit Markus' Diktat besser beurteilt als im Kontext mit Danielas Diktat.***

4. *Kommen Lehrer und Lehrerinnen bei Korrektur und Bewertung in etwa zu gleichen Ergebnissen?* Bei der Korrektur trauten die Studierenden Lehrern wie Lehrerinnen zu, die Fehlerzahl richtig zu identifizieren, da es als möglich angesehen wurde, im Zweifelsfall unter Zuhilfenahme eines Dudens die Anzahl der Fehler objektiv zu bestimmen. Insofern lautete die Hypothese: ***4a) Bei der Fehlerzahl gibt es keine Geschlechtsunterschiede.*** Bei der Einschätzung der Geschlechtsunterschiede in Bezug auf die Bewertung der Diktate waren eventuell die Ergebnisse von Newton (1942), Edminston (1943) und Carter (1952, 1953) zu berücksichtigen. Alle Untersuchungen zeigten übereinstimmend, dass die Lehrerinnen nicht nur generell mildere Noten erteilten, sondern dass sie außerdem den Mädchen deutlich bessere Noten gaben als den Jungen, während das bei den Lehrern im umgekehrten Fall weniger ausgeprägt zu beobachten war. Insofern war die Hypothese: ***4b) Lehrerinnen geben mildere Zensuren als Lehrer.***

5. *Beeinflusst das Dienstalter die Präzision der Korrektur und die Milde der Bewertung des Diktats?* Da es sich bei der Korrektur um die Anwendung leicht zu beherrschender Messaspekte der Rechtschreibleistung handelt, glaubten wir nicht an deren Beeinflussung durch das Dienstalter. Bei dem Aspekt der Bewertung allerdings könnte man sich vorstellen, dass das Dienstalter doch eine Rolle spielt. Junge Lehrkräfte dürften noch gute Erinnerungen daran haben, wie man sich selbst fühlte, wenn man schlechte Noten bekam, wie viel Ängste dadurch ausgelöst wurden, wie sehr man sich dadurch auch demotiviert fühlte. Bei jungen Gymnasiallehrern war eine entsprechende Urteilstendenz bei mündlichen Prüfungen festgestellt worden (Birkel 1978). Das könnte auch bei jungen Grund- und Hauptschullehrkräften der Fall sein. Daher wurde die Hypothese aufgestellt: ***5) Junge Lehrkräfte geben zumindest mildere Beurteilungen ab als sog. „alte Hasen".***

6. *Hängt das Ergebnis von der Tatsache ab, dass die Lehrer selbst im Fach Deutsch unterrichten oder nicht?* Im Bereich der Grund- und Hauptschule ist „fachfremder" Unterricht immer noch eher die Regel als die Ausnahme. Auch Lehrkräfte, die das Fach Deutsch nicht studiert haben, müssen vor allem dann, wenn sie als Klassenlehrer fungieren, oft auch das Fach Deutsch unterrichten, obwohl sie es in ihrer Ausbildung nicht studiert haben. Den Lehrkräften wird die grundsätzliche Fähigkeit, sich in ein unbekanntes Fachgebiet einzuarbeiten und Unterricht sinnvoll zu gestalten, zugeschrieben. Unter der Voraussetzung, dass diese Zuschreibung gerechtfertigt ist, lautet die entsprechende Hypothese: ***6) Die Tatsache, dass die Lehrkräfte selbst Deutsch als Fach unterrichten oder nicht, wirkt sich nicht auf die Diktatbeurteilung aus.***

7. *Spielt es eine Rolle, ob die Diktate von Lehrern beurteilt wurden, die schwerpunktmäßig an der Grund- oder an der Hauptschule unterrichten?* Eigentlich konnte

man hier keine wirklich begründete Hypothese erstellen, weil es zu dieser Fragestellung vermutlich noch keine Untersuchung gibt. Unsere Erwartung ging aber in die Richtung, dass eventuell Lehrer, die vornehmlich in der Grundschule arbeiten, die Neigung haben könnten, Schülerleistungen positiver zu bewerten, weil bei Grundschülern mehr intrinsische Motivation über die Leistungsrückmeldungen geweckt werden sollte. Die volle „Härte des schulischen Alltags" erreicht die Schüler in der Regel erst im Hauptschulalter. Da das alles aber nur Vermutungen sind, lautet unsere Hypothese vorsichtshalber: ***7) Grund- und Hauptschullehrer unterscheiden sich nicht in ihren Bewertungen.***

8. Für wie schwierig wird der Diktattext gehalten und inwieweit beeinflusst die Schwierigkeitseinschätzung die Beurteilung? Einhellige Überzeugung aller Seminarteilnehmer war, dass den Schülern eigentlich eine Art Bonus gewährt werden müsste, wenn die Lehrer den Diktattext als sehr schwer einstufen würden. Die Ursache für die schlechten Leistungen könnte dann eher dem Lehrer angelastet werden, der den Text entwickelte, als dem Schüler, der sich mit der richtigen Schreibung schwer tat. Da es aber auch hier keinerlei empirische Belege für eine solche Urteiltendenz gibt, einigte man sich auf eine Nullhypothese: ***8) Die Schwierigkeitseinschätzung hat keinen Einfluss auf die Diktatbeurteilung.***

4. Untersuchungsdesign

Bei der Beurteilung von Diktatleistungen gibt es u.U. eine Vielzahl möglicher Einflussfaktoren. Dass die gezeigte Rechtschreibleistung ein solcher Faktor ist, darf unterstellt werden. Dass das aber der alleinige Einflussfaktor ist, darf eher bezweifelt werden, auch wenn meist so getan wird, als sei dem so. Vielfältige empirische Befunde zeigen aber immer wieder, dass leistungsfremde Faktoren in die Beurteilung eingehen und dadurch die Objektivität der Beurteilungen herabmindern (Ingenkamp 1971-1995[9]; Krapp 1984; Kühn 1983; Tent, Fingerhut & Langfeldt 1976; Krapp 1973). Solche Faktoren können in der Person des Lehrers, der Person des Schülers, der Gestaltung der Leistungssituation (Aufgabe) oder in der Situation der Schule liegen.

Im varianzanalytischen Design für das Diktat von Paul wurden die vier Faktoren „Kontext", „Geschlecht der Lehrer", „Dienstaltersgruppe" und „Schulart" berücksichtigt. Da bis auf das Dienstalter (3 Stufen) alle anderen Faktoren jeweils 2-stufig waren, ergaben sich insgesamt 24 Teilstichproben, die allerdings nicht immer gleich stark besetzt waren. Bei den Schülern Markus und Daniela reduzierte sich das Design auf drei Faktoren mit jeweils 12 Zellen.

Version 1: Diktate Paul + Markus												Version 2: Diktate Paul + Daniela											
Lehrer						Lehrerin						Lehrer						Lehrerin					
Jung		Mittel		Alt		Jung		Mittel		Alt		Jung		Mittel		Alt		Jung		Mittel		Alt	
GS	HS	GS	HS	GS	HS	GS	HS	GS	HS	GS	HS	GS	HS	GS	HS	GS	HS	GS	HS	GS	HS	GS	HS
Anzahl Fehler																							
Zensur																							

Abbildung 1

Der Einfluss der Schwierigkeitseinschätzung auf die Schülerleistungen und der Frage, ob Deutsch als Fach unterrichtet wird, soll bei allen drei Schülern in weiteren Einweg-Varianzanalysen geprüft werden.

5. Stichprobe

In die Untersuchung gingen die Diktatbewertungen von insgesamt 415 Lehrkräften aus den Regierungsbezirken Nord-, Südwürttemberg und Südbaden des Bundeslandes Baden-Württemberg ein. Dass der Regierungsbezirk Nordbaden nicht berücksichtigt wurde, lag an der Zusammensetzung der Studierenden der PH Weingarten, die ganz überwiegend aus dem eigenen Regierungsbezirk Südwürttemberg stammen. Einige Studierende stammten aber auch aus Nordwürttemberg und Südbaden. Die große Anzahl von Lehrern dürfte aber dafür gesorgt haben, dass eine vermutlich gute Repräsentativität der Stichprobe gewährleistet war, auch wenn diese nicht explizit überprüfbar war.

Da nicht immer alle Lehrkräfte alle Angaben machten, vor allem auch bei den personenbezogenen Daten, reduzierte sich die Stichprobe hier und da etwas. Von den 415 Lehrkräften beurteilten 194 das Diktat von Paul im Kontext mit Markus, 221 im Kontext mit Daniela. 142 Lehrer und 266 Lehrerinnen gaben ihr Geschlecht an. 7 Lehrkräfte verweigerten die Angabe. Die Dienstaltersstreuung reichte bei den Lehrkräften von 1 bis zu 43 Dienstjahren. 9 Lehrer machten keine Angaben zu ihrem Dienstalter.

Die Einteilung in Dienstaltersgruppen erfolgte so, dass in etwa ähnlich große Gruppen entstanden. In der Gruppe der „jungen Lehrer“ wurden die Lehrkräfte mit bis zu 10 Dienstjahren zusammengefasst. Lehrer mit 11 bis 28 Dienstjahren bildeten die „mittelalte“ Gruppe und die mit 29 bis 43 Dienstjahren die „alte“ Gruppe. Da einige Lehrer ihr Dienstalter nicht angaben, entfielen auf diese drei Gruppen jeweils 176, 98 bzw. 132 Lehrkräfte. In der relativ großen Gruppe der jungen und alten Lehrer spiegelt sich die Einstellungspraxis der Landesregierung in Baden-Württemberg wider, die lange Zeit kaum neue Lehrer einstellte und erst in den letzten Jahren wieder vermehrt auch junge Lehrer eingestellt hat. Hinzu kam sicher auch die größere Bereitschaft jüngerer Lehrkräfte, sich an einer solchen Untersuchung zu beteiligen.

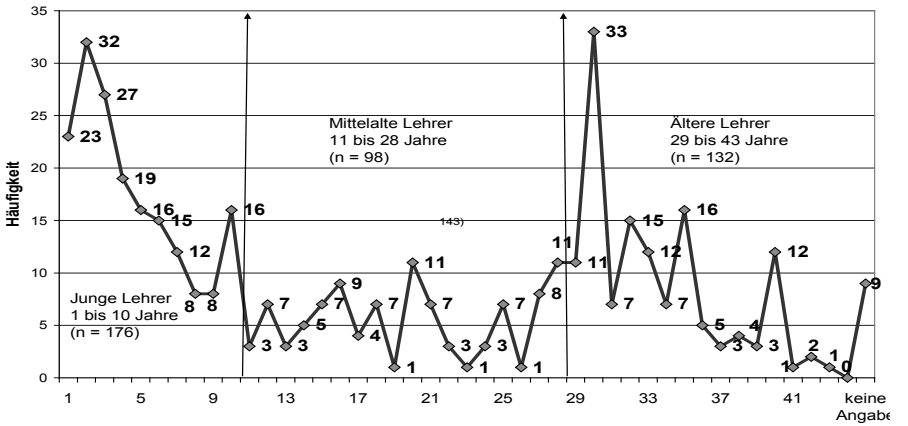

Abbildung 2: Dienstaltersverteilung der Lehrer

Da bis auf 32 fast alle Lehrer über Unterrichtserfahrung im Fach Deutsch verfügten und nur 8 dazu keine Angaben machten, lässt sich feststellen, dass 375 Lehrkräfte entweder das Fach Deutsch zum Zeitpunkt der Untersuchung oder in der Vergangenheit unterrichteten.

Bei der Frage, ob die Lehrkräfte vorwiegend an Grund- oder Hauptschulen unterrichteten, gaben 270 an, vorwiegend an der Grundschule zu arbeiten. 136 arbeiteten vorwiegend an der Hauptschule und 9 hatten vergessen, hier ihr Kreuz zu setzen. Damit hatten etwa doppelt so viele Grundschullehrkräfte teilgenommen wie Hauptschullehrkräfte.

6. Ergebnisse

6.1 Übereinstimmung der Lehrkräfte

Identifizierte Fehler

Gerade bei der Anzahl der identifizierten Fehler war die Erwartung sehr hoch angesetzt, denn die Frage, wie viele Fehler die drei Schüler in ihren Diktaten hatten, sollte sich doch recht eindeutig beantworten lassen. Hier gab es nun die erste Überraschung, denn die Streubreiten der identifizierten Fehler fielen wesentlich größer aus als erwartet.

Aus Abbildung 3 ist zu erkennen, dass die Anzahl der identifizierten Fehler beim Diktat von Daniela eine Streubreite über 10 Fehler zeigte. Das Minimum lag bei ganzen 2 Fehlern, das Maximum bei immerhin 11 Fehlern. Im Durchschnitt wurden 6,74 Fehler angestrichen.

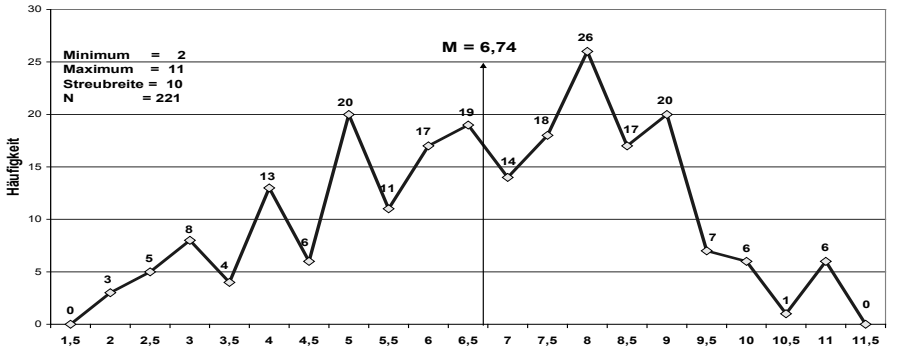

Abbildung 3: Fehlerzahl beim Diktat von Daniela

Insgesamt 389 Lehrkräfte hatten Pauls Diktat begutachtet. Es lässt sich erkennen, dass minimal 10 Fehler gefunden wurden, maximal aber 28, damit also fast dreimal so viele Fehler. Im Schnitt wurden 19 Fehler angestrichen. Der Anteil der Lehrkräfte im Bereich Mittelwert ± 1 ganzer Fehler liegt mit 40,6% recht hoch. Es erstaunt aber doch, dass drei Lehrkräfte nur jeweils 10 Fehler fanden, während andererseits eine Lehrkraft immerhin 28 Fehler fand.

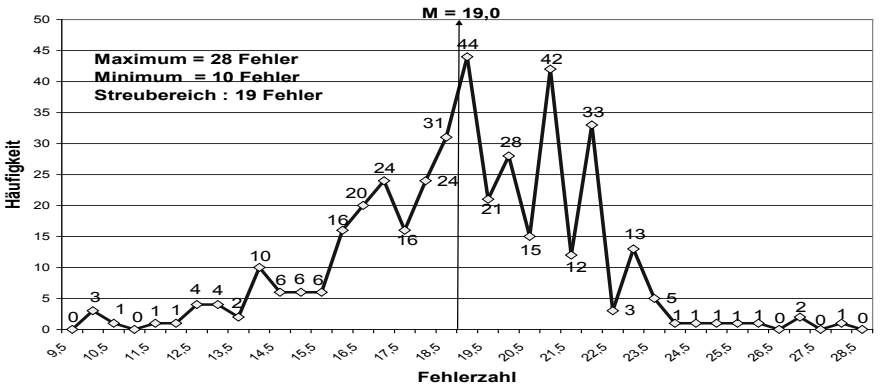

Abbildung 4: Fehlerzahl beim Diktat von Paul

185 Lehrer hatten das Diktat von Markus begutachtet. Hier gab es noch größere Unterschiede in der Anzahl der identifizierten Fehler. Aus Abbildung 5 wird deutlich, dass die geringste Fehlerzahl bei 34,5, die höchste bei 74,5 und damit mehr als doppelt so hoch lag. Die Streubreite reichte somit über 41 Fehler. Im Schnitt waren gut 60 Fehler gefunden worden. Fasst man wiederum die Anteile der Lehrkräfte zusammen, die in die Fehlerkategorien oberhalb und unterhalb des Mittelwerts fallen, so kann man sagen, dass 37,3% der Lehrer zwischen 58 und 63,5 Fehler gefunden hatten. Unsere *Hypothese*, dass sich die Anzahl der jeweils gefundenen Fehler

nicht wesentlich unterscheiden dürfte, *kann unter diesen Umständen wohl kaum weiter vertreten werden*.

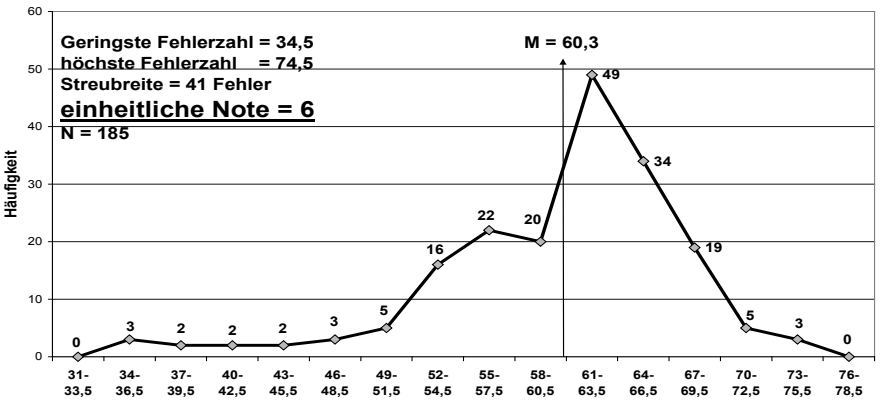

Abbildung 5: Fehlerzahl beim Diktat von Markus

Zensuren

Wenn schon die Zahl der identifizierten Fehler sich so deutlich unterschied, dann musste sich das auch - entgegen unserer Erwartung - in den erteilten Zensuren niederschlagen.

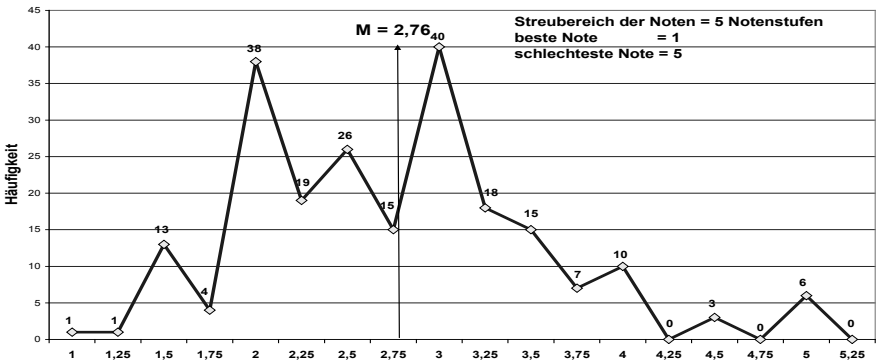

Abbildung 6: Notenverteilung beim Diktat von Daniela

In Abbildung 6 ist die Verteilung der Zensuren dargestellt, die die Lehrkräfte für Danielas Diktat für angemessen hielten. Was wirklich niemand erwartet hatte, war die große Streubreite der Zensuren, die von der glatten 1 bis 5 reichte. Zwei Lehrkräfte gaben die Note 1 und 1- , während immerhin sechs Lehrkräfte der Auffassung waren, dass die Diktatleistung mit der Note 5 zu bewerten sei. Der Noten-

durchschnitt aller Lehrkräfte lag bei 2,76, was etwa der Note 3+ entspricht. Damit lag in dieser Untersuchung die durchschnittliche Beurteilung deutlich schlechter als die im Original erteilte Note, denn der Klassenlehrer hatte Daniela eine 2 gegeben. Trotz der erheblichen Beurteilungsunterschiede kann festgestellt werden, dass 62% der Lehrer mit ihren Noten im Bereich zwischen der 2 und der 3 lagen. Dass in den Randbereichen die Noten so weit auseinandergingen, verblüffte.

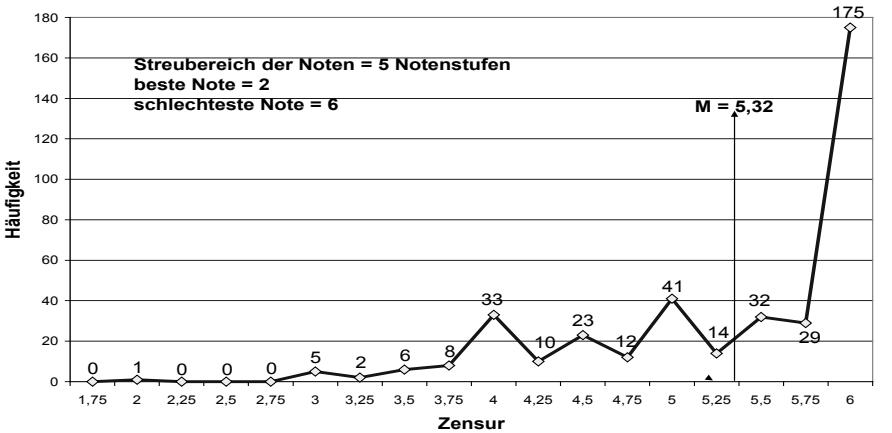

Abbildung 7: Notenverteilung beim Diktat von Paul

Beim Diktat von Paul registriert man erneut, dass die Streubreite der Zensuren fünf Notenstufen beträgt und von der 2 bis zur 6 reicht. Paul bekam im Schnitt die Note 5,32, was in etwa der 5- entspricht. Erneut bewegen sich die Lehrkräfte zu fast ¾ im Bereich der Noten 5 und 6. Immerhin ein Viertel aller Lehrer gibt aber positivere Zensuren, bei denen die eine Lehrkraft, die die Note 2 gibt, sicher als Ausreißer nicht ganz ernst genommen werden darf. Wir erinnern uns, dass der Klassenlehrer hier die Note 4 gegeben hatte. Erneut liegt er damit deutlich besser als der Schnitt der Lehrkräfte. Damit erweist sich unsere Hypothese 2 erneut als nicht haltbar. Die Originalbeurteilung und das Notenmittel dieser Untersuchung weichen im Vergleich zum Diktat von Daniela noch deutlicher voneinander ab.

Auf eine grafische Darstellung der Notenverteilung von Markus kann verzichtet werden, denn sein Diktat war nun wirklich so fehlerbehaftet, dass sich alle Lehrkräfte – einschließlich des Klassenlehrers – einig darin waren, dass die hier gezeigte Leistung absolut ungenügend war.

Unsere Hypothese 1b, dass sich die Lehrer bezüglich der zu erteilenden Zensuren relativ einig sein dürften, lässt sich also nur für die absolut schlechteste Leistung aufrecht erhalten. Bei den etwas oder deutlich besseren Leistungen gehen dagegen die Bewertungen erstaunlich weit auseinander. Einen gewissen subjektiven Spielraum war man ja bereit, den Lehrern zuzugestehen, aber dass die Streubreiten dann über fünf Notenstufen reichten, war überraschend.

6.2. Ergebnisse der Varianzanalysen

Um die Haupteffekte dieser Untersuchung und deren Wechselwirkungen auf die Identifikation von Fehlern und die Beurteilung in Form von Zensuren überprüfen zu können, wurde für den Schüler Paul eine 4-faktorielle Varianzanalyse mit den Faktoren A = Version, B = Geschlecht der Lehrkräfte, C = Schulart und D = Dienstaltersgruppe gerechnet. Bei den Schülern Daniela und Markus reduzierte sich das auf eine jeweils 3-faktorielle Varianzanalyse, bei der der Faktor „Version" wegfiel. Die Analyseergebnisse sind den Tabellen im Anhang 1 zu entnehmen, der auf der Internetseite von „Didaktik Deutsch" eingestellt ist.

Schaut man sich dort die Varianzanalysetabellen für den Schüler Paul an (Tab. 1), dann fällt auf: Bei der Fehlerzahl werden etwa 15% der Varianz durch die experimentellen Faktoren und deren Interaktionen erklärt, gut 85% der Varianz bleiben als Fehlervarianz ungeklärt. Bei den Zensuren (Tab.2) beläuft sich die geklärte Varianz auf immerhin 31% und nur 69% bleiben als Fehlervarianz ungeklärt. Da aber unsere experimentellen Faktoren alle als außerhalb der zu beurteilenden Leistung liegende zu betrachten sind, wird sofort klar, dass die Leistungsbeurteilung viel anfälliger für subjektive Beeinflussungen ist als die Leistungsmessung, nämlich die Identifikation der Fehler.

Beim Schüler Markus musste nur die Varianzanalyse für die identifizierte Fehlerzahl berechnet werden (Tab. 3), denn bei der Note waren sich alle Lehrer einig. Alle erteilten seiner Leistung die Note 6, es gab also keine Varianz, die man hätte analysieren können! Bei der Schülerin Daniela, die ja ein recht gutes Diktat abgeliefert hatte, mussten wieder beide Varianzanalysen berechnet werden: für die identifizierte Fehlerzahl und für die Zensur (Tab. 4+5).

Kontexteffekt bei Schüler Paul

Nur Pauls Diktat war mit verschiedenen Referenzleistungen zu beurteilen gewesen. In der Version 1 ging die sehr schlechte Diktatleistung von Markus voraus, in der Version 2 die relativ gute von Daniela. Hier war es nun spannend nachzuschauen, ob der Kontext, in dem das Diktat von Paul zu beurteilen war, Auswirkungen auf die Auswertung und die Beurteilung haben würde.

Im Gegensatz zu unserer Hypothese 3a zeigt sich bei der Auswertung von Pauls Diktat ein sehr signifikanter Kontexteffekt (*F=7,17; df=1/384; p<.01*). Mit der Referenzleistung von Markus (Version 1) wurden in Pauls Diktat im Schnitt 18,2 Fehler angestrichen, während im Kontext mit Daniela (Version 2) 19,8 Fehler markiert wurden. Es wurden also im Mittel gut eineinhalb Fehler weniger gefunden, wenn Pauls Diktat nach dem von Markus auszuwerten war. Mit 1,59% war allerdings die Varianzklärung hier nicht sehr hoch.

Der Kontexteffekt verstärkt sich bei der Analyse der Beurteilungsunterschiede noch einmal (*F=23,58; df=1/384, p<.001*). Die Irrtumswahrscheinlichkeit verringert sich also auf weniger als 1‰. Damit kann man sagen, dass der Kontexteffekt deutlicher zutage tritt, wenn im Bereich der Beurteilung die Subjektivität im Vergleich zur Auswertung zunimmt. Die Varianzklärung steigt hier auf 4,23% an. Die Relevanz

dieses Unterschiedes ist somit als deutlich größer anzusehen im Vergleich zu den Unterschieden bei der Fehlerfindung.

Mit Markus' Diktat als Referenzleistung (Version 1) bekommt Paul die Note 4,99, während er bei der Referenzleistung von Daniela die 5,35 erhält. Die Notendifferenz beträgt zwar „nur" 0,36 Notenstufen, aber man muss bedenken, dass es sich dabei um Mittelwertsunterschiede handelt. Damit kann die Hypothese 3b als bestätigt angesehen werden. Paul erhält bessere Beurteilungen wenn die Referenzleistung sehr schlecht war (Markus, Version 1) und schlechtere, wenn die Referenzleistung gut war (Daniela, Version 2).

Version 1	18,2	4,99
Version 2	19,8	5,35

Tabelle 1: Durchschnitt der gefundenen Fehler und der Noten bei Paul

Geschlecht der Lehrer

Bei unserer Hypothese 4 gingen wir davon aus, dass sich bei der Identifikation der Fehler Lehrer und Lehrerinnen nicht, wohl aber bei der Beurteilung unterscheiden könnten. Das soll nun geprüft werden.

	Fehler	Note
Lehrer	17,9	5,02
Lehrerin	19,3	5,32

Tabelle 2: Identifizierte Fehler und Noten bei Paul

Wenden wir uns zuerst dem Schüler Paul zu, der von allen Lehrkräften zu beurteilen war. Aus Tab. 2 kann abgelesen werden, dass sich die Anzahl gefundener Fehler bei Lehrern und Lehrerinnen signifikant unterschied (*F=22,56; df=1/384; p<.001; Eff%=5,0*). Damit ist unsere Hypothese zumindest für diesen Schüler zurückzuweisen. Die Lehrerinnen haben im Schnitt 1.3 Fehler mehr identifiziert als die Lehrer.

Als nächstes wäre dann zu fragen, welchen Einfluss das auf die Beurteilung des Diktats hatte. Analog zu den mehr gefundenen Fehlern gaben die Lehrerinnen auch im Schnitt die schlechteren Zensuren (*F=15,25; df=1/384; p<.001; Eff%=2,74*). Die Differenz der Durchschnittsnoten liegt bei 0,3 Notenstufen. Damit wäre Hypothese 4b zurückzuweisen. Entgegen den bisher berichteten Untersuchungsergebnissen zu den Urteilstendenzen von Lehrern und Lehrerinnen urteilen hier die Lehrerinnen strenger als ihre männlichen Kollegen. Da der Faktor Geschlecht auch noch bei einigen Interaktionen mit anderen Faktoren signifikant wurde, lässt sich der deutliche Geschlechtseffekt auf eine bestimmte Gruppe von Lehrern eingrenzen. Es stellt sich nämlich heraus, dass es vor allem die männlichen Grundschullehrer sind, die so milde Zensuren erteilen.

In gleicher Weise müssen dann auch die Geschlechtsunterschiede bei den beiden anderen Schülern überprüft werden, die allerdings jeweils nur von etwa der Hälfte der Lehrer beurteilt worden waren.

Wenden wir uns jetzt der Schülerin Daniela zu, die ja ein vergleichsweise gutes Diktat geschrieben hatte. Es wird sofort deutlich, dass bei den Fehlerzahlen kein Geschlechtseffekt auftritt (*F=0,02; n.s.*). Entsprechend unserer Hypothese identifizieren die Lehrkräfte diesmal unabhängig von ihrem Geschlecht annähernd gleich viele Fehler (♂ = 6,78; ♀ = 6,80). Bei den Zensuren sehen die Verhältnisse ähnlich aus (*F=0,19; n.s.*). Wieder geben männliche und weibliche Lehrkräfte fast identische Noten (♂ = 2,75; ♀ = 2,79). Zumindest bei der relativ guten Diktatleistung wird somit unsere Hypothese bestätigt.

Bei Markus ist von vornherein schon klar, dass es bei den Zensuren keinen Geschlechtsunterschied geben kann, denn alle Lehrer waren sich in der Bewertung ja einig. Ein Diktat mit so vielen Fehlern konnte nur mit einer 6 beurteilt werden.

	Fehler
Lehrer	58,8
Lehrerin	61,1

Tabelle 3: Fehler bei Markus

Spannender ist da schon die Frage, ob sich Lehrer und Lehrerinnen bei der Anzahl gefundener Fehler unterscheiden würden. Gehen die Lehrerinnen ähnlich wie bei Paul mit größerer Genauigkeit auf Fehlersuche? Der Haupteffekt beim Schüler Markus wird signifikant. Lehrerinnen streichen, ähnlich wie bereits bei Paul, mehr Fehler an als Lehrer (*F=4,99; df=1/173; p<.05; Eff%=2,5*).Die Differenz beträgt diesmal 2,3 Fehler im Schnitt. Sie ist zwar größer als bei Paul, aber wegen der generell höheren Fehlerzahl ist dieser Unterschied „nur“ auf dem 5%-Niveau signifikant.

Versucht man nun einen Gesamtüberblick über die gefundenen Geschlechtsabhängigkeiten bei den drei Diktaten zu geben, so lässt sich sagen, dass in zwei Fällen die Unterschiede bei den gefundenen Fehlern zugunsten der Lehrerinnen ausfielen, einmal war der Unterschied nicht signifikant. Es scheint also so, dass die Lehrerinnen mit größerer Akribie an die Fehlersuche gegangen sind, vor allem wenn auch wirklich viele Fehler vorlagen. Die vertretene Nullhypothese ließ sich allerdings nur für das relativ gute Diktat von Daniela aufrechterhalten.

In Bezug auf die Beurteilungstendenzen erscheint das Bild etwas anders. Hier kann festgestellt werden, dass die Beurteilungen sich im Prinzip an den gefundenen Fehlern orientieren. Wenn wie bei Paul die Lehrerinnen mehr Fehler finden, geben sie auch entsprechend schlechtere Zensuren. Wichtig erscheint der Hinweis, dass trotzdem die Notenvergabe hier stärker als geschlechtsabhängig anzusehen ist als die Identifikation von Fehlern, denn die Signifikanz steigt deutlich an. Finden Lehrer und Lehrerinnen aber wie bei Daniela fast gleich viele Fehler, so unterscheiden sich auch die gegebenen Zensuren höchstens zufällig.

Fazit: Unsere Hypothese bezüglich der Bewertung der Diktate erweist sich als zutreffend, wenngleich mit umgekehrtem Vorzeichen. Lehrerinnen geben hier nicht die milderen, sondern im Gegenteil sogar die strengeren Zensuren.

Dienstalter

Die Studierenden gingen im Seminar davon aus, dass sich die Lehrkräfte unterschiedlichen Dienstalters nicht bei der Identifikation der Fehler unterscheiden würden, wohl aber bei der Bewertung. Dieser Sachverhalt soll bei den drei beteiligten Schülern überprüft werden.

	Fehler	Note
junge Lehrer	18,7	5,05
mittelalte Lehrer	18,4	5,18
alte Lehrer	18,5	5,29

Tabelle 4: Fehlerzahl und Zensuren bei Paul

Beginnen wir wieder mit dem Diktat von Paul, das ja praktisch von allen Lehrkräften korrigiert und bewertet wurde. Entsprechend unseren Erwartungen zeigte sich, dass bei den Korrekturen kein Dienstalterseffekt auftrat (*F=0,38; n.s.*). Die Lehrkräfte identifizierten in allen drei Altersklassen in etwa gleich viele Fehler. Bei den Zensuren dagegen ergab sich ein signifikanter Haupteffekt (*F=3,3; df=2/384; p<.05; Eff.%=1,18*). Erwartungsgemäß stiegen die Durchschnittsnoten in den Alterskategorien von 5,05 bei den jungen Lehrkräften fast linear auf 5,29 bei den alten Lehrkräften an. Der Unterschied zu den mittelalten Lehrkräften war nach beiden Seiten nicht signifikant. Die jungen Lehrer haben also nur milder geurteilt als die alten.
Betrachtet man die Verhältnisse bei Daniela, dann ist festzustellen, dass sowohl bei den Fehlern als auch bei den Zensuren kein signifikanter Haupteffekt auftritt.

Unterrichtserfahrung im Fach Deutsch

Bei der Überprüfung der Frage, wie viele der beteiligten Lehrkräfte überhaupt das Fach Deutsch selbst unterrichteten, ließ sich festhalten: Nur 32 Lehrkräfte hatten *keine* Unterrichtserfahrung im Fach Deutsch, 358 dagegen sehr wohl. Eventuelle Unterschiede wurden mit einer Einweg-Varianzanalyse überprüft.

Beim Schüler Paul lässt sich feststellen, dass hypothesengetreu keine Unterschiede bei den Korrekturergebnissen aufgetreten sind. Erfahrene wie auch weniger erfahrene Lehrkräfte kommen praktisch zum gleichen Ergebnis.

	Note
Deutsch nein	4,88
Deutsch ja	5,35

Tabelle 5: Notenschnitt bei Paul

Bei der Notenvergabe dagegen urteilten die Lehrkräfte ohne eigene Deutscherfahrungen deutlich milder (*F=29,22; df=1/398; p< .001*). Dem relativ schwachen Rechtschreiber Paul kam die mangelnde Erfahrung der Lehrkräfte also zugute. Die Mittelwertsdifferenz lag bei 0,47 Notenstufen. Lehrer ohne Unterrichtserfahrung im Fach Deutsch urteilten um fast eine halbe Notenstufe imSchnitt milder als die Kollegen mit entsprechender Unterrichtserfahrung. Die Varianzklärung lag ebenfalls mit knapp 10% in einem Bereich, den man als sehr relevant bezeichnen kann.

Als nächstes folgt die Überprüfung bei Schülerin Daniela. 20 Lehrkräfte hatten keine Unterrichtserfahrung im Fach Deutsch, 201 hatten sie. Hier unterscheiden sich die Lehrkräfte mit und ohne Unterrichtserfahrung im Fach Deutsch sowohl bei der Zahl der gefundenen Fehler als auch bei der Beurteilung. Der Haupteffekt bei der Fehlerzahl war signifikant (*F=6,26; df=1/220; p<.01*). Lehrer ohne eigene Unterrichtserfahrung im Fach Deutsch fanden signifikant mehr Fehler. Die Differenz beträgt 0,67 Fehler, was bei der geringeren Fehlerzahl von Daniela sehr bedeutsam ist. Die Varianzklärung von 2,78% kann schon als relevant angesehen werden.

	Fehler	Note
Deutsch nein	6,82	2,78
Deutsch ja	6,15	2,51

Tabelle 6: Fehlerzahl und Notenschnitt bei Daniela

Noch deutlicher werden die Unterschiede bei der Notenvergabe (*F=7,86; df=1/220; p<.01*). Der F-Wert steigt etwas an, die Varianzklärung steigt auf 3,46%. Auch die Differenz von 0,27 Notenstufen ist demnach sehr bedeutsam. Da bei der relativ geringen Fehlerzahl viele Lehrkräfte *mit* Deutscherfahrung hier weniger Fehler anstrichen, gaben diese auch entsprechend mildere Noten. Die mangelnde Deutscherfahrung der Lehrer kam der relativ guten Rechtschreiberin *nicht* zugute! Das war bei Paul noch ganz anders. Da das Diktat von Daniela immer im Kontext mit Pauls Diktat zu beurteilen war, kann man vermuten, dass sich der Referenzeffekt in der Form bemerkbar machte, dass die weniger erfahrenen Lehrer den schlechten Schüler besser und die bessere Schülerin schlechter beurteilt haben. Bei der besseren Schülerin kann deren Rechtfertigung allerdings darin gesehen werden, dass sie dort auch mehr Fehler fanden und ihre Zensur daran anzupassen hatten.

	Fehler
Deutsch nein	63,9
Deutsch ja	60,1

Tabelle 7: Fehlerzahl bei Markus

Schließlich bleibt beim Schüler Markus noch die Überprüfung eventueller Abhängigkeiten der Ergebnisse von der Deutscherfahrung der Lehrer übrig. Hier ist allerdings wieder nur die Überprüfung der Fehlerzahlen möglich, weil als Zensur ja einheitlich die Note 6 gewählt worden war. Keine Deutscherfahrung hatten hier nur 11 der insgesamt 185 Lehrkräfte. Ähnlich wie bei Daniela konnte Markus *nicht* von

der Unerfahrenheit der Lehrkräfte profitieren. Gerade diese Lehrer strichen im Schnitt 3,8 Fehler mehr an. Das war auf dem 1‰-Niveau signifikant (*F=17,23; df=1/184; p<.001*) und bei einer Varianzklärung von 8,61% auch ziemlich relevant.

Betrachtet man die Effekte bei den drei Schülern insgesamt, kann man feststellen, dass bei der Identifikation der Fehler zweimal die Lehrkräfte ohne Deutscherfahrung mehr Fehler angestrichen hatten, und einmal gab es keinen Unterschied. Damit scheint die Frage, ob die Lehrer Deutsch selbst schon unterrichtet hatten oder nicht, die Korrektur doch stärker beeinflusst zu haben. Unsere Hypothese der Unterschiedslosigkeit lässt sich so vermutlich nicht aufrechterhalten.

Im Hinblick auf die Bewertungen ergab sich ebenfalls zweimal ein deutlicher Unterschied. Unsere Hypothese der Unterschiedslosigkeit, die wir eigentlich nur mangels vertretbarer Belege so erstellt hatten, erscheint auch hier vermutlich als eher nicht vertretbar. Da aber einmal Paul im positiven Sinne von den geringeren Deutscherfahrungen profitiert, bei Daniela aber ein entgegengesetzter Effekt sichtbar wird, könnte man auch argumentieren, dass sich diese Effekte sozusagen gegenseitig aufheben, was wiederum für die Beibehaltung unserer Nullhypothese spräche.

Schulart

Bei Paul kann man konstatieren, dass der Haupteffekt des Faktors Schulart bei der Fehlerzahl als abhängiger Variable signifikant wird (*F=9,65; df=1/384; p<.01; Eff%=2,14*). Die Anzahl gefundener Fehler unterscheidet sich bei Grund- und Hauptschullehrern deutlich. Grundschullehrer strichen 0,8 Fehler im Schnitt weniger an. Das steht zunächst einmal im Widerspruch zu unserer Nullhypothese.

	Fehler	Note
Hauptschule	18,9	5,47
Grundschule	18,1	4,87

Tabelle 8: Fehlerzahl und Notenschnitt bei Paul

Beim Notenschnitt allerdings waren die Unterschiede mit einer Differenz von 0,6 Notenstufen beträchtlich. Dieser Unterschied ist hoch signifikant (*F=63,27; df=1/384; p<.001; Eff.%=11,35*). Mit einer Varianzklärung von über 11% ist er auch als sehr relevant zu bezeichnen. Halten wir fest: Bei wenig mehr gefundenen Fehlern geben die Hauptschullehrer bei Paul deutlich strengere Noten als die Grundschullehrer!

	Fehler	Note
Hauptschule	7,2	2,97
Grundschule	6,4	2,58

Tabelle 9: Fehlerzahl und Notenschnitt bei Daniela

In gleicher Weise sind wieder die Angaben zu Daniela, der besseren Rechtschreiberin, zu überprüfen. Bei der leistungsstarken Daniela stellen sich die Verhältnisse ähnlich dar wie bei Paul. Erneut finden die Hauptschullehrer mehr

Fehler als die Grundschullehrer (*F=11,67; df=1/205; p<.001; Eff.%=5,09*). Auch bei der Bewertung der Rechtschreibleistung kehren die Verhältnisse wieder. Erneut erteilen die Hauptschullehrer deutlich strengere Noten (*F=15,01; df=1/205, p<.001; Eff.%=6,51*), und erneut steigt die Varianzklärung an. Die Mittelwertsdifferenz liegt hier bei 0,39 Notenstufen. Damit lässt sich die Hypothese der Unterschiedslosigkeit nicht aufrechterhalten. Die Zugehörigkeit zur Schulart scheint als außerhalb der eigentlichen Leistung liegender Einflussfaktor bei der Auswertung und Beurteilung der Diktate eine Rolle zu spielen.

Schwierigkeitseinschätzung durch die Lehrkräfte

Bei der Schwierigkeitseinschätzung wagten die Seminarteilnehmer keine Erwartungen zu formulieren, die sich auf die Korrekturergebnisse und die Zensuren bezogen. Insofern zog man sich auf die Nullhypothese zurück und meinte, dass alles andere als eine Unabhängigkeit der Auswertung und Bewertung des Diktats von der Schwierigkeitseinschätzung eine Überraschung wäre. Andererseits schwang aber doch auch die Erwartung mit, dass die Lehrkräfte den Schülern einen Bonus gewähren könnten, wenn sie die Verantwortlichkeit für die hohe Schwierigkeit dem Klassenlehrer zuschreiben würden.

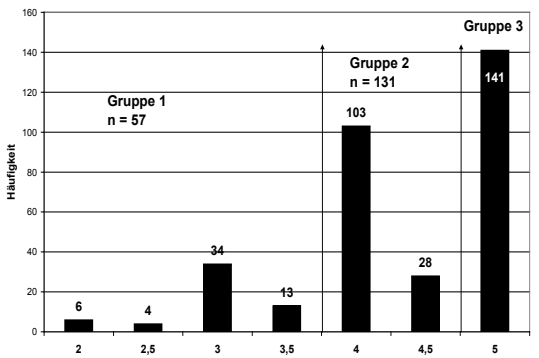

Abbildung 8: Schwierigkeitseinschätzung

Die Lehrer hatten auf einer fünfstufigen Skala, die von „1 = sehr leicht“ bis „5 = sehr schwer“ reichte, ihre Schwierigkeitseinschätzung frei abzugeben. Leider haben relativ viele Lehrkräfte hier keine Einschätzung abgeben mögen, weshalb die Stichprobengrößen hier deutlich geringer ausfielen. Die Verteilung der Schwierigkeitsbeurteilungen ist aus Abbildung 8 zu ersehen.

Um den Einfluss der Einschätzung auf Auswertung und Bewertung des Diktats überprüfen zu können, wurden drei Gruppen gebildet. In Gruppe 1 wurden die Lehrer aufgenommen, die meinten, das Diktat sei mittelschwer (Werte 2 bis 3,5; 57 Personen). Gruppe 2 bildeten die Lehrer, die das Diktat für eher schwer hielten (Werte 4 und 4,5; 131 Personen) und als Gruppe 3 fungierten alle Lehrer, die der Meinung waren, das Diktat sei sehr schwer gewesen (Wert 5; 141 Personen).

Nach Auffassung der Lehrkräfte war das Diktat insgesamt gesehen doch als eher schwer anzusehen. Die große Mehrheit betrachtete das Diktat als eher schwer (4) und sogar als sehr schwer (5). Kein Lehrer war der Meinung, dass dieses Diktat sehr leicht (1 oder 1,5) gewesen sei. Nur sechs meinten, es sei eher leicht (2). Der Durchschnitt der Einschätzungen lag beim Wert 4,3!

Die Einwegvarianzanalysen wurden wieder getrennt nach den drei Diktaten vorgenommen. Wenden wir uns zunächst dem Einfluss der Einschätzung auf das Diktat von Paul zu.

Sowohl bei der Korrektur als auch bei der Bewertung sind die Unterschiede zwischen den gebildeten Gruppen bei Paul relativ gering. Da die Effekte nicht signifikant wurden, muss man in beiden Fällen davon ausgehen, dass die vorhandenen Differenzen wohl eher dem Zufall zuzuschreiben sind als einem Bonus, den Paul vielleicht dadurch hätte erhalten haben können, dass die Verantwortung für die Schwierigkeit des Diktats dem Klassenlehrer angelastet wurde. Trotz der Insignifikanz scheinen die Mittelwertsunterschiede aber in die erwartete Richtung zu gehen. Lehrkräfte, die das Diktat für sehr schwer hielten, strichen geringfügig weniger Fehler an und gaben einen Hauch bessere Noten. Aber wie gesagt, die Unterschiede liegen hier im Bereich der Zufälligkeit!

Aufgabe	Fehlerzahl	Noten
mittelschwer	7,48	2,98
eher schwer	6,61	2,78
sehr schwer	6,70	2,68

Tabelle 10: Fehler und Noten bei Daniela

In gleicher Weise wurden dann die Korrekturen und Bewertungen für Danielas Diktat geprüft. Bei Daniela wirkt sich die Schwierigkeitseinschätzung signifikant sowohl bei der Fehlerzahl ($F=4,25;\ df=2/215;\ p<.05$) als auch beim Notenschnitt ($F=30,06;\ df=2/215,\ p<.01$) aus. Wenn das Diktat von der Schwierigkeit her als angemessen eingeschätzt wurde (2 – 3,5), wurden im Schnitt 0,87 Fehler mehr angestrichen, als wenn das Diktat als eher schwer (4 – 4,5) eingeschätzt wurde. Wurde es sogar als sehr schwer eingeschätzt (5), fanden die Lehrkräfte nur unwesentlich mehr Fehler als wenn es als eher schwer eingestuft worden war. Man kann also sagen, dass Daniela von der Schwierigkeitseinschätzung in der erwarteten Richtung profitierte.

Da verwundert es nicht mehr, wenn sich dieser Effekt auch bei der Bewertung von Danielas Diktat aufzeigen lässt. Sowohl bei der Einschätzung „eher schwer“ als auch bei „sehr schwer“ gaben die Lehrkräfte signifikant mildere Beurteilungen ab, auch wenn die Differenz im Notenschnitt bei maximal 0,3 Notenstufen liegt. Das, was bei Paul als nicht signifikante Tendenz sich abzeichnet, tritt bei Daniela deutlich zutage: Wird das Diktat als eher oder sehr schwer betrachtet, gibt man der Schülerin einen gewissen Bonus und beurteilt die Leistung besser. Damit kann die Hypothese 8 nicht weiter aufrechterhalten werden. Unsere in diesem Zusammenhang geäußerten Erwartungen haben sich eher bestätigt. Erneut tritt der Effekt auf, dass die Beur-

teilung stärker als die Leistungsmessung von Faktoren beeinflusst wird, die nicht direkt in die Zensur eingehen müssten, hier also die Schwierigkeitseinschätzung des Diktats.

7. Bewertung der Ergebnisse

Versucht man sich die Vielfalt der Teilergebnisse noch einmal zu vergegenwärtigen und die wichtigsten Dinge zu benennen, dann kann man auf folgende Erkenntnisse aus dieser Studie verweisen:

1. Die Erwartung, dass die ***Korrektur*** von Diktaten objektiv geschehen und damit zu übereinstimmenden Ergebnissen führen könnte, erweist sich als trügerisch. Die minimale Fehlerzahl muss mit dem Faktor 2,5 bis 3 versehen werden, um auf die Maximalzahl zu kommen. Eine solche Streuung bei den gefundenen Fehlern erstaunt, weil in der Regel davon ausgegangen wird, dass anhand eines Dudens zweifelsfrei entschieden werden kann, ob ein Wort richtig oder falsch geschrieben wurde. Diese Erwartung wurde vermutlich auch erfüllt, aber die individuellen Zielsetzungen der Lehrkräfte bei der Korrektur bestimmen auch mit, welche „lässlichen" Fehler als solche angekreidet werden oder nicht. Hier ist vor allem zu nennen, wie mit ausgelassenen i-Punkten, Umlauttüpfeln oder t-Strichen verfahren wird. Werden diese Fehler überhaupt angestrichen? Falls ja, werden sie als halbe oder ganze Fehler gewertet? Werden Fehler bei der Interpunktion überhaupt mitbewertet und falls ja, wie stark? Allein diese Frage macht klar, dass die ***Kriterien***, nach denen ein Diktat korrigiert wird, für die auswertenden Lehrkräfte ***nicht definiert*** und damit ihrem subjektiven Ermessen überlassen sind. Es könnte dann größere Einigkeit herrschen bezüglich der Frage, ob nur die Fehler bewertet werden, die dem behandelten Rechtschreibproblem angehören, oder alle anderen Fehler auch, wenn es entsprechende, leicht anwendbare Handlungsanweisungen gäbe. Dabei ist auch klar, dass eine Handlungsanweisung die didaktische Freiheit des Lehrers einschränken kann und darum nicht unbedingt willkommen sein muss.

In jüngerer Vergangenheit wurden Anleitungen für die Auswertung von Diktaten publiziert und in der Fachdidaktik propagiert. Solche Analysesysteme sind z. B. DoRA (Löffler & Meyer-Schepers, 1992), AFRA (Herné & Naumann, 2002) oder OLFA (Thomé & Thomé, 2004). Sie bieten differenzierte Anweisungen zur Auswertung von Rechtschreibleistungen mit der Zielsetzung, für die Schüler spezielle Rückmeldungen zu ermöglichen, wo deren Problemschwerpunkte liegen (s. dazu auch Herné, 2006). So kann dann auch die Lehrkraft individuelle Fördermöglichkeiten für Schüler bereitstellen. Ein Problem liegt nur in der Frage, inwieweit die Lehrkräfte solche Systeme kennen und anwenden können!

Ein weiteres Problem, das die Objektivität der Korrektur beeinträchtigt, liegt in der – vorsichtig ausgedrückt – „individuellen" Art der Schüler zu schreiben. Schüler schaffen es immer wieder, Buchstaben so zu schreiben, dass die Lehrkraft u.U. viele Interpretationsmöglichkeiten hat. So kann ein „a" leicht sehr ähnlich geschrieben werden wie ein „o". Unterstellt man nun dem Schüler, dass er den Buchstaben geschrieben hat, der an dieser Stelle richtig war, erkennt man das Wort als richtig an. Glaubt man dem Schüler das nicht, oder ist die Lehrkraft einfach jemand, der auf

gute Lesbarkeit und schöne Schrift großen Wert legt, wird sie hier einen Fehler markieren. Ein konkretes Beispiel aus dem Diktat von Paul sei angeführt (s. Anhang auf der Internetseite von „Didaktik Deutsch"): Es kamen viele Wörter mit den Buchstaben „p" oder „pp" vor. Offensichtlich hatte Paul die Angewohnheit, die Unterlängen des Buchstaben „p" sehr kurz zu halten. Dadurch sah dieser Buchstabe sehr ähnlich aus wie der Buchstabe „r". Wenn jetzt das Wort „Treppe" zu schreiben war, konnte man es auch als „Trerre" lesen. Manch eine Lehrkraft hat darum dieses Wort immer wieder als Fehler angestrichen, andere dagegen entschieden, dass es sich dabei um eine Eigenheit der individuellen Schrift des Schülers handele, und akzeptierten diese Wörter als richtig. Damit ist aber klar, dass dem subjektiven Ermessen des Lehrers hier wieder eine Tür geöffnet wurde.

Als weiteres, weithin bekanntes Problem war die Frage beobachtbar, wie man mit Wörtern umzugehen habe, die mehrfach im Text vorkommen und mehrfach falsch geschrieben werden. Wird das Wort als solches einmal als Fehler gewertet oder wird es immer wieder als Fehler gewertet? Bei einzelnen Lehrkräften wurde deutlich, dass sie die verschärfte Version für angemessen hielten. Sogar der Fall einer Lehrkraft wurde beobachtet, die auch in einem Wort mehrere Fehler anstrich und wertete. Dieses Vorgehen kann dann als richtig akzeptiert werden, wenn es darum geht, bestimmte Fehlerschwerpunkte für den Schüler zu identifizieren, um gezielte Übungen anschließen zu können. Dass auf diese Weise aber sehr unterschiedliche Fehlerzahlen auftraten, ist unmittelbar nachvollziehbar! Hier hätte man als Experimentator den Lehrern Hinweise geben können, wie man sich selbst die Korrektur wünscht. Das ist aber nicht geschehen, weil ja die individuellen Korrekturstrategien der Lehrerschaft zum Tragen kommen sollten.

Es lässt sich also festhalten, dass ***die Korrektur eines Diktats bei Weitem nicht zu so objektiven Ergebnissen*** führt, wie viele Lehrkräfte glauben. Die Anzahl der identifizierten Fehler kann sich von Lehrer zu Lehrer stark unterscheiden. Auf diesen Tatbestand haben fachdidaktische Publikationen schon häufiger verwiesen (vgl. etwa Menzel 1997).

2. Bei der ***Beurteilung*** der Diktatleistung kommen traditionell noch viel mehr subjektive Überzeugungen und Motive ins Spiel. Soll der Schüler durch die Note motiviert werden, seine Anstrengungen zu erhöhen, oder möchte man dem Schüler nur einfach ein Erfolgserlebnis gönnen? Damit verbunden ist immer auch die Entscheidung der Lehrkraft darüber, welchen Bewertungsmaßstab sie verwenden will. Kommt es eher darauf an, die Position eines Schülers im Vergleich zur sozialen Gruppe aller Schüler in der Klasse darzustellen, den individuellen Lernfortschritt zu bewerten oder das Erreichen bestimmter unterrichtlicher Ziele zu attestieren? Soll eine Note einen ungehörigen Schüler disziplinieren? Für wie bedeutsam hält die Lehrkraft die Fähigkeit, normgerecht schreiben zu können? Wie überzeugt ist die Lehrkraft von der Notwendigkeit, Diktate schreiben zu sollen? Diese Einstellungen könnten Auswirkungen haben darauf, wie „scharf" oder „milde" die Lehrkraft urteilt.

Im zweistufigen Prozess der Leistungsbeurteilung kann man den Aspekt der Messung als einer Objektivierung eher zugänglich ansehen, während beim eigentli-

chen Beurteilungsprozess zwangsläufig wieder stärker subjektive Einflüsse wirksam werden. Letzteres ist durchaus nicht als Vorwurf zu werten, denn die erzieherische Aufgabe des Lehrers macht es manchmal ja geradezu nötig, einen gewissen subjektiven Spielraum nutzen zu können. Insofern muss eben fast zwangsläufig auch die Beurteilung weniger „objektiv" ausfallen als die Messung der Leistung. Bisher unbeantwortet bleibt allerdings die Frage, wie viel Subjektivität man der Lehrkraft zugestehen möchte, die diesen Prozess verantwortungsbewusst zu bewältigen hat. Hilfestellungen seitens der Fachdidaktik werden der Lehrkraft nur sehr bedingt angeboten. In den gängigen Fachdidaktiken wird auf das Problem der Leistungsbeurteilung und seine Lösungsmöglichkeiten nicht so konkret eingegangen, dass Lehrkräfte hier wirkliche Orientierung bei der Überführung individueller Rechtschreibleistungen in Zensuren finden können. Anleitung zur Aus- und Bewertung von Rechtschreibleistungen bietet bisher nur die Oldenburger Fehleranalyse (OLFA) von Thomé & Thomé (2004). Anhand eines differenzierten Analyseschemas ermöglicht es die Auswertung der Fehler in den Kategorien „Vor- bis protoalphabetisch", „Alphabetisch" und „Orthographisch" und bietet Formeln an, nach denen sich Kompetenzwerte berechnen lassen.

In der vorliegenden Untersuchung wurde bei der Frage, wie die Lehrer zu ihrer Bewertung der Diktatleistungen gekommen sind, häufig auf eine andere Formel verwiesen, die die Umrechnung der identifizierten Fehler in eine „objektive", weil vom Lehrer unabhängige Note ermöglicht. Leider wusste niemand mehr, woher diese Formel stammt. Sie lautete:

$$Note = 1 + \frac{Fehlerzahl}{GesamtzahlWörter} * 50$$

Natürlich wird der Umrechnungsprozess von Fehlerzahlen in Noten durch solche Formeln ein Stück weit objektiviert, aber gleichzeitig akzeptiert die Lehrkraft bestimmte Setzungen, die damit verbunden sind. Leider sind sich viele Lehrkräfte dieser impliziten Festlegungen nicht bewusst. Eine dieser Setzungen ist, dass man unterstellt, dass ab einem Fehleranteil von 8% der insgesamt geschriebenen Wörter die Note 5 erteilt wird. Dort liegt also der „cut-off-point" zwischen einer noch akzeptablen und einer als nicht mehr hinreichend empfundenen Leistung. Wenn 10% der Wörter oder mehr falsch geschrieben werden, würde die Note „6 = ungenügend" zu erteilen sein. Wer garantiert aber dafür, dass diese Festlegungen sinnvoll sind? Würde man die Addition des Wertes 1 weglassen, läge besagter „cut-off-point" statt bei 8% der Wörter bei 10%.

Ein anderer hinterfragenswerter Punkt wäre die Festlegung auf die Multiplikation des Quotienten aus Fehlerzahl und Gesamtzahl der Wörter mit dem Faktor 50. Im Prinzip böte diese Formel hier eine Möglichkeit, die Schwierigkeit eines Diktattextes entsprechend zu berücksichtigen. Es mag ja sein, dass bei einem durchschnittlich schweren Diktat der „cut-off-point" von 8% der Gesamtwortzahl sinnvoll ist. Wenn allerdings ein Diktattext einmal deutlich schwieriger ist, dann könnte man den entsprechenden Faktor reduzieren. Bisher sind aber weder Regelungen bekannt, wie man einerseits die Schwierigkeit eines Diktattextes objektiv bestimmt, noch wie man dann diesen Schwierigkeitsindex in den zu verwendenden Faktor umrechnet,

um den Schülern damit gerecht zu werden. Eine weitere unhinterfragte Setzung dieser Formel wäre also, dass das Diktat von durchschnittlicher Schwierigkeit wäre. Besonders bei dem in dieser Untersuchung verwendeten Diktat hätte die Schwierigkeit des Textes durchaus zu einer entsprechenden Herabsetzung dieses Faktors führen können, denn eine Vielzahl von Lehrkräften attestierten diesem Text einen hohen Schwierigkeitsgrad. Offenbar war sich der Klassenlehrer dieser Schwierigkeit deutlicher bewusst, denn seine Beurteilungen für die Schüler Paul und Daniela waren milder ausgefallen als die durchschnittlichen Beurteilungen aller beteiligten Lehrkräfte.

Dann bleibt bei der vorgeschlagenen Formel noch die Frage nach der Sinnhaftigkeit einer linearen Beziehung zwischen der Fehlerzahl und der Note bei der Bewertung eines Diktats offen. Muss die Differenz der Fehlerzahlen von Notenstufe zu Notenstufe gleich sein? Die oben angeführte Formel geht einfach davon aus. Trotzdem wäre zu fragen, ob nicht mit abnehmender Leistung die Differenzwerte auch größer werden könnten, ähnlich wie Hiller (2004) das darstellt. Wäre es nicht auch möglich, vielleicht sogar sinnvoll, wenn die Fehlerdifferenz zwischen den Noten „ausreichend“ und „mangelhaft“ größer wäre als zwischen den Noten „sehr gut“ und „gut“? In solch einem Fall hätte man es mit einer kurvilinearen Beziehung zwischen Fehlerzahl und Note zu tun. Pädagogisch ließe sich auch ein solches Verfahren rechtfertigen. An dieser Stelle soll keine Entscheidung für das eine oder andere Vorgehen gefällt werden. Das bleibt Aufgabe der Fachdidaktik und der verantwortungsbewussten Lehrkraft. Aus allem bisher Dargelegten geht aber klar hervor, dass die Beurteilung einer Diktatleistung auf der Notenskala noch weit mehr Platz für subjektive Einflüsse seitens der beurteilenden Lehrkraft bietet, als das in der Öffentlichkeit gemeinhin gesehen wird.

Fazit in dieser Hinsicht könnte sein: Beurteilungsprozesse sind immer stärkeren subjektiven Einflüssen ausgesetzt als die eigentlichen Messprozesse. In die Beurteilung bringt sich jeder Mensch mit seiner Erfahrung und seinen Motiven und Einstellungen ein. Das führt zu unterschiedlichen Bewertungen einzelner Teilaspekte und damit auch zu unterschiedlichen Beuteilungen insgesamt. Da auch Lehrer nur Menschen sind, können auch sie sich solchen Einflüssen nicht entziehen. Sie sollten sich nur dieser Einflüsse bewusst sein und sie nach außen hin auch zugeben können. Beurteilungen sind immer hoch komplexe und nicht immer nur rational gesteuerte psychische Prozesse von Menschen.

3. Betrachten wir noch einmal die Einflussfaktoren auf die Beurteilung des Diktats, das dieser Untersuchung zugrunde lag.

- Der **Kontext** einer Beurteilung bestimmt immer mit, wie man eine bestimmte Leistung beurteilt. Kein Mensch verfügt über „absolute Maßstäbe“ zur Leistungsbeurteilung ähnlich dem „absoluten“ Gehör. Da bei unserer Untersuchung ein Diktat zu beurteilen war ohne Kenntnis der bisherigen Diktatleistungen der Schüler, war verständlich, dass jeder Lehrer nach Referenzpunkten sucht, die es ihm erleichtern, eine Leistung zumindest nach dem meist verwendeten, dem sozialen Bewertungsmaßstab zu bewerten. Insofern war es logisch nachvollziehbar, dass die unterschiedlich angebotenen Referenzleistungen die Beurteilung von Pauls Diktat beeinflussten.

Dass auch darüber hinaus jeder Lehrer die Möglichkeiten nutzen kann, seine Beurteilungen je nach Referenzrahmen anzupassen, ist weithin angewendete Praxis. Was macht der Lehrer, der bemerkt, dass sein Diktat zu schwer war? Er kann zumindest seinen Bewertungsmaßstab so anpassen, dass er trotzdem allen Notenstufen eine Leistung zuordnet.

- Das **Geschlecht des Lehrers** spielt bei der Beurteilung der Diktatleistungen eine Rolle. Lehrer fanden weniger Fehler und beurteilten infolgedessen auch die Leistungen positiver als die Lehrerinnen. Dieser Effekt trat besonders deutlich beim Diktat des Schülers Paul zutage. Die signifikanten Interaktionen legen dann aber eine differenziertere Sicht zu diesem Haupteffekt nahe. Es war in dieser Untersuchung vor allem die Gruppe Grundschullehrer bei der Version 1 (Diktate Markus & Paul), die dem Diktat von Paul eine besonders milde Beurteilung zukommen ließen. Sie waren es vor allem, die sich bei der Beurteilung an der Referenzleistung orientierten, während bei allen anderen Gruppen keine bedeutsamen Unterschiede auftraten.

- Auch das **Dienstalter** kann die Auswertungs- und Beurteilungsprozesse beim Diktat beeinflussen. Beim Schüler Paul findet sich eine fast lineare Beziehung zwischen den Dienstaltersgruppen und den Zensuren. Erwartungsgemäß geben die jungen Lehrer mildere Noten als die älteren. Obwohl das Dienstalter an weiteren Interaktionen beteiligt war, ergaben sich keine Differenzierungen bezüglich des Dienstalterseffekts, wenn man einmal absieht von der Tatsache, dass die jüngeren Lehrkräfte bei Markus weniger Fehler angestrichen hatten als die älteren.

- Bei der Frage, ob die **Unterrichtserfahrung im Fach Deutsch** Auswertung und Beurteilung des Diktats beeinflussen, ergaben sich widersprüchliche Befunde. Die relativ schwache Rechtschreibleistung von Paul wird von den Lehrkräften ohne eigene Erfahrung im Fach Deutsch milder beurteilt. Die relativ gute Rechtschreiberin Daniela kann dagegen nicht von diesem Mangel an Erfahrung profitieren. Bei ihr werden im Gegenteil von diesen Lehrkräften mehr Fehler angestrichen und konsequenterweise dann auch strengere Zensuren erteilt. Aussagen über den Einfluss der eigenen Unterrichtserfahrung auf die Beurteilung eines Diktats können also vermutlich nur im Zusammenhang mit weiteren, eventuell moderierenden Faktoren getroffen werden, zumal die Teilstichprobe der Lehrkräfte ohne eigene Unterrichtserfahrung im Fach Deutsch recht klein war.

- Auch die **Schulart**, an der die Lehrkräfte überwiegend eingesetzt sind, kann als bedeutsamer Faktor angesehen werden. Grundschullehrer streichen durchgängig weniger Fehler an und geben dementsprechend auch mildere Noten. Besonders deutlich wird das bei den jungen Lehrkräften, während die Unterschiede sich bei den mittelalten und älteren Lehrkräften eher wieder verwischen.

- Der Eindruck von der **Schwierigkeit des Diktattextes** wirkt sich ebenfalls auf Auswertung und Beurteilung der Diktate aus. Wird der Text für schwer gehalten, profitieren die Schüler eher davon. Das war besonders deutlich zu erkennen bei der Schülerin Daniela. Auch wenn die entsprechenden Effekte bei den beiden anderen

Schülern nicht signifikant waren, so lagen doch die Kennwerte jeweils in der gleichen erwarteten Richtung.

Bleibt zum Abschluss noch festzuhalten, dass selbst so vermeintlich objektive Tätigkeiten wie das Identifizieren von Rechtschreibfehlern bei Weitem nicht so objektiv leistbar sind, wie es immer wieder vermutet wird. Neben den vielen Problemen im Hinblick auf die nicht gekannten oder nicht definierten Kriterien der Auswertung und Beurteilung kommen auch noch die außerhalb der eigentlichen Rechtschreibleistung liegenden Faktoren hinzu, die solche Zensuren verzerren können. Dass die Beurteilung von Aufsätzen eine schwierige Aufgabe für die Lehrer darstellt, wird heute allseits anerkannt. Dass aber auch Diktate und Mathematikarbeiten vergleichbaren Problemen unterliegen, scheint die Lehrerschaft wie die Öffentlichkeit eher zu verblüffen. Die Objektivität der Leistungsbeurteilung ist und bleibt ein unerfüllbarer Wusch, auch bei Klassenarbeiten, denen man ein höheres Maß an Objektivität zugetraut hätte. Lehrer täten gut daran, sich das immer wieder einmal ins Gedächtnis zu rufen, wenn sie weitreichende Entscheidungen über die schulische Laufbahn einzelner Schüler zu treffen haben. Das trifft besonders dann zu, wenn solche Entscheidungen aufgrund einzelner (Prüfungs-)Arbeiten gefällt werden müssen.

Die Ergebnisse dieser Untersuchung könnten ein weiteres Argument gegen die weitreichende Anwendung von Diktaten geliefert haben oder eben für die Erstellung gewisser Regeln, die bei der Fehlerwertung zu beachten sind. Ob allerdings alternative Formen der Lernkontrolle im Bereich der Rechtschreibung zu objektiveren Ergebnissen führen, darf wohl bezweifelt werden. Das Geschäft der Leistungsmessung im Bereich der Rechtschreibung bleibt schwierig!

Literatur

Adrion, D. (1984). Rechtschreiben. In J. Baurmann & O. Hoppe (Hrsg.). *Handbuch für Deutschlehrer*. Kohlhammer: Stuttgart.

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

Birkel, P. (1984). Beurteilung mündlicher Prüfungsleistungen. In K. Heller (Hrsg.). *Leistungsdiagnostik in der Schule*. Huber: Bern, S. 229-236.

Birkel, P. (2003). Aufsatzbeurteilung – ein altes Problem neu untersucht. *Didaktik Deutsch, 9*, Heft 15, S. 46-53.

Birkel, P. (2005). Beurteilungsübereinstimmung bei Mathematikarbeiten? *Journal für Mathematik-Didaktik, 26*, S. 28-51.

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

Birkel, P. & Stammet, C. (2008). Die Entwicklung der Rechtschreibfähigkeit von Schülern der Grund- und Hauptschule aus der Sicht einer Neueichung des WRT. In W. Schneider, H. Marx & M. Hasselhorn (Hrsg.). *Diagnostik von Rechtschreibleistungen und –kompetenzen*. Jahrbuch der Pädagogisch-psychologischen Diagnostik. Tests und Trends N.F. Band 6, S. 61-91. Hogrefe: Göttingen.

Brinkmann, E. (2004). Schreiben nach Diktat oder selbstständig Rechtschreibung lernen? *Grundschule*, Heft 1, S. 11-13.

Carter, R.S. (1952). How invalid are marks assigned by teachers? *The Journal of Educational Psychology, 43,* No. 4, S. 216-228. Gekürzte Wiedergabe unter dem Titel: Wie gültig sind die durch Lehrer erteilten Zensuren? In K. Ingenkamp (Hrsg.). *Die Fragwürdigkeit der Zensurengebung.* Beltz: Weinheim, 1995[9], S. 148-158.

Carter, R.S. (1953). Non-intellectual variables involved in teacher's marks. *Journal of Educational Research,* S. 81-95.

Edminston, R.W. (1943). Do teachers show partiality toward boys or girls? *Peabody Journal of Education, 20*, S. 234-238.

Fix, M. (1991). Möglichkeiten des differenzierenden Umgangs mit Diktaten. *Praxis Deutsch, 18*, H. 108 „Differenzieren und Individualisieren", S. 47-52.

Fix, M. (1994). *Geschichte und Praxis des Diktats im Rechtschreibunterricht*. Lang: Frankfurt/M.

Fix, M. (2004). Funktionen des Diktats und Diktatkritik. *Grundschule*, Heft1, S. 8-10.

Heller, K. & Rosemann, B. (1974). *Planung und Auswertung empirischer Untersuchungen.* Klett: Stuttgart.

Herné, K.L. & Naumann, C.L. (2002). *Aachener Förderdiagnostische Rechtschreibfehler-Analyse (AFRA)*. Alfa Zentaurus: Aachen.

Herné, K.L. (2006). Rechtschreibtests. In U. Bredel et al. (Hrsg.). *Didaktik der deutschen Sprache. Band 2.* (S. 883-897) Schöningh UTB 8236: Paderborn.

Hiller, E.M. (2004). Das Dilemma mit den Diktaten. *Die Grundschulzeitschrift*, Heft 180, S. 38-42.

Ingenkamp, K. (Hrsg.) (1971, 1995[9]). *Die Fragwürdigkeit der Zensurengebung.* Beltz: Weinheim.

Ingenkamp, K. & Lissmann, U. (1985, 2005[5]). *Lehrbuch der Pädagogischen Diagnostik.* Beltz UTB: Weinheim.

Kleiter, E. (1988). *Lehrbuch der Statistik in KMSS. Band 11: Überblick und niedrig-komplexe Verfahren*. Deutscher Studienverlag: Weinheim.

Kleiter, E. (1990). *Lehrbuch der Statistik in KMSS. Band 1/2: Niedrig-komplexe Verfahren.* Deutscher Studienverlag: Weinheim.

Kleiter, E. (2004). *KMSS-8. Kleiter-Microcomputer-Statistik-System*. Kiel

Korn, W. (2005). *Formen der Leistungserhebung im Fach Deutsch*. Auer: Donauwörth.

Krapp, A. (1984). Forschungsergebnisse zur Bedingungsstruktur der Schulleistung. In K. Heller (Hrsg.). *Leistungsdiagnostik in der Schule*. Huber: Bern, S. 46-62.

Krapp, A. (1973). *Bedingungen des Schulerfolgs*. Oldenbourg: München.

Kühn, R. (1983). *Bedingungen für Schulerfolg*. Hogrefe: Göttingen.

Leßmann, B. (2004). DIKTATE – ohne Ende? Schritte zur endgültigen Verabschiedung des traditionellen Diktates. *Grundschulunterricht,* Heft 4, S. 33-39.

Löffler, I. & Meyer-Schepers, U. (1992). *DoRA: Dortmunder Rechtschreibfehler-Analyse zur Ermittlung des Rechtschreibstatus rechtschreibschwacher Kinder*. Dortmund.

Menzel, W. (1997). Diktieren und Diktiertes aufschreiben. *Praxis Deutsch*, Heft 142, S. 15-26.

Newton, R.F. (1942). Do men teachers grade higher than women teachers? *School and Society, 56*, S. 72.

Richter, S. & Brügelmann, H. (Hrsg.) (1994). *Mädchen lernen ANDERS als Jungen. Geschlechtsspezifische Unterschiede beim Schriftspracherwerb.* DGLS-Reihe „Lesen und Schreiben". Libelle: CH-Lengwil.

Spitta, G. (1976). Wozu überhaupt diktate? und: Rechtschreibföderalismus: Diktate – zwang oder selbstverordnete praxis? *Die Grundschule, 8*, S.478-484.

Stolla, G. (2001). Mit Fehlerzählen ist es nicht getan. Praxis Deutsch, Heft 144, S. 30-32.

Tent, L., Fingerhut, W. & Langfeldt, H.-P. (1976). *Quellen des Lehrerurteils*. Beltz: Weinheim.

Tent, L. & Birkel, P. (2009, im Druck). Zensuren. In: D.H. Rost (Hrsg.). *Handwörterbuch Pädagogische Psychologie.* 3., völlig überarbeitete Auflage. Beltz PVU: Weinheim.

Thiel, O. & Valtin, R. (2002). Eine Zwei ist eine Drei ist eine Vier. In: R. Valtin: *Was ist ein gutes Zeugnis?* (S. 67-76) Juventa: München.

Thomé, G. & Thomé, D. (2004). *Oldenburger Fehleranalyse OLFA. Version 2.0.* Igel Verlag Wissenschaft: Oldenburg.

Zillig, M. (1928). Einstellung und Aussage. *Zeitschrift für Psychologie, 106*, S. 58-106.

Anschrift des Verfassers:
Dr. Peter Birkel, Doggenriedstr. 18, 88250 Weingarten, Tel.: 0751/52007, E-Mail: birkel@ph-weingarten.de