

Einfluss des Leistungsniveaus einer Schulklasse auf die Benotung am Ende des 3. Schuljahres

Anke Treutlein, Jeanette Roos und Hermann Schöler

Schulklassen unterscheiden sich in ihrem Leistungsniveau. Für Schüler/-innen hat der unterschiedliche Kontext Auswirkungen: Nicht nur das Selbstkonzept wird von der Referenzgruppe beeinflusst (Big-Fish-Little-Pond-Effekt), auch Lehrkräfte können bei der Notengebung solchen Referenzgruppeneffekten unterliegen.

An 1'256 Kindern wurde der Einfluss der Referenzgruppe auf die Benotung am Ende der 3. Klasse überprüft. Unter Berücksichtigung des Klassenleistungsniveaus wurde die Leistung der Kinder in Lese- und Rechtschreibtests zur Benotung durch die Lehrkraft in Beziehung gesetzt.

Die Ergebnisse zeigen nicht den erwarteten Referenzgruppeneffekt: Der individuelle Leistungsstand hat größeren Einfluss auf die Benotung als das Klassenleistungsniveau. Mögliche Ursachen sowie das grundsätzliche Problem, wie stark das Klassenleistungsniveau in den Noten berücksichtigt werden muss, werden diskutiert.

Einleitung

Bereits im Jahr 1969 verwies Ingenkamp darauf, dass sich Schulklassen eines Jahrgangs in der Variationsbreite ihrer Leistungsfähigkeit unterscheiden können, obgleich das so genannte Jahrgangsklassensystem auf der Annahme einer weitgehenden Entsprechung von Alter, Entwicklung und Leistungsfähigkeit beruht. Werden objektive Leistungsmaße herangezogen, kann ein Kind bei gleicher Testleistung in Abhängigkeit vom Leistungsniveau in der einen Klasse zu den leistungsstärksten, in einer anderen Klasse zu den leistungsschwächsten Kindern der Klasse gehören (Scheib, Schöler, Fehrenbach, Roos & Zöllner, 2005; Trautwein & Baeriswyl, 2007). Der jeweils unterschiedliche Klassenleistungskontext in Relation zur eigenen Position im Leistungsgefüge der Klasse bleibt für die Selbst- und Fremdbewertung der Kinder nicht ohne Folgen. In Abhängigkeit von der Leistungsstärke der Klasse fallen aufgrund sozialer Vergleichsprozesse zum Beispiel Selbstbewertungen hinsichtlich der eigenen Fähigkeiten unterschiedlich aus. Dieser aus der Forschung zum Einfluss der Referenzgruppe auf

das Selbstkonzept bekannte Big-Fish-Little-Pond-Effekt (Marsh, 1987) besagt, dass ein Kind, das in einer (in Bezug auf das Abschneiden in standardisierten Tests) insgesamt leistungsschwachen Klasse zu den besten Kindern gehört, zu einer positiveren Einschätzung der eigenen Fähigkeiten gelangt als ein (test)leistungsgleiches Kind, das in einer insgesamt sehr leistungsstarken Klasse zu den schlechteren Schüler/-innen zählt.

Referenzgruppeneffekte und Fremdbewertung

Referenzgruppeneffekte wirken sich nicht nur auf das Fähigkeitsselbstkonzept von Schüler/-innen aus. Wie Trautwein, Lüdtke, Marsh, Köller und Baumert (2006) zeigen konnten, besteht zudem ein systematischer Zusammenhang zwischen den von Lehrkräften erteilten Schulnoten und dem mittleren (durch standardisierte Leistungstests erhobenen) Leistungsniveau von Klassen: Je höher das Leistungsniveau der Klasse, desto schlechter wird eine gleiche Testleistung benotet. Soziale Vergleichsprozesse beeinflussen also auch die Fremdbewertung der Leistung, da auch Lehrkräfte den jeweiligen Klassenkontext als Bezugssystem bei der Leistungsbeurteilung ihrer Schüler/-innen heranziehen. Aufgrund dieses Vergleichs wird Schülern/-innen in ihrer Klasse ein Rangplatz zugewiesen. Dieser Rangplatz sagt zwar etwas über die relative Position von Schülern/-innen in der Leistungshierarchie ihrer Klasse aus, gibt aber keine Auskunft über die tatsächlich erbrachte Leistung. Die dabei berücksichtigte so genannte soziale Bezugsnorm ist nicht wie bei normierten Verfahren klassenübergreifend, sondern hat nur klassenintern Bestand und lässt damit das tatsächliche Leistungsniveau der Klasse bei der Bewertung individueller Leistungen außer Acht (Rheinberg, 2001; Schrader & Helmke, 2001). Dies führt dazu, dass Noten die Leistungsunterschiede zwischen den Kindern einer Klasse zwar gut wiedergeben (Ingenkamp, 1969; Schrader & Helmke, 2001), allerdings über verschiedene Klassen hinweg nicht vergleichbar sind (Tent, 2001). Gleiche Testleistungen korrespondieren in den einzelnen Klassen mit ganz unterschiedlichen Noten. Zeugnisnoten geben demzufolge die Leistungsunterschiede zwischen den Klassen nicht wieder (Ingenkamp, 1969; Thiel & Valtin, 2002) und sind daher kein geeignetes Maß für den tatsächlichen Leistungsstand eines Kindes.

Mit Hilfe von Mehrebenenmodellen, die der hierarchischen Datenstruktur von Leistungen einzelner Schüler/-innen innerhalb verschiedener Klassen gerecht werden, konnte gezeigt werden, dass dieselbe Testleistung in leistungsstarken Klassen negativer beurteilt wird als in leistungsschwachen Klassen (Trautwein & Baeriswyl, 2007). Der dabei beobachtete Effekt entspricht dem Big-Fish-Little-Pond-Effekt und fällt sehr bedeutsam aus. Auch die Übergangsempfehlungen am Ende der Grundschulzeit sind abhängig vom Leistungsniveau der Klasse: Bei gleicher Testleistung ist in leistungsstarken Klassen die Chance eine Gymnasialempfehlung zu bekommen geringer als in leistungsschwachen Klassen (Tiedemann & Billmann-Mahecha, 2007; Trautwein & Baeriswyl, 2007). Die angeführten Studien zeigen, dass nicht nur die Selbstbewer-

tung im Sinne eines Big-Fish-Little-Pond-Effekts vom jeweiligen Umfeld beeinflusst wird, sondern auch die Fremdbewertung der Leistung vom Kontext abhängig ist. Bei einer Vorgabe von Mindeststandards, wie dies in schulischen Curricula der Fall ist, sollte die Benotung von Schüler/-innenleistungen idealerweise anhand einer von Außen gesetzten kriterialen Bezugsnorm erfolgen (Rheinberg, 2001). Bei lernzielorientierten Vergleichen existiert ein Leistungskriterium oder -ziel, und es werden eher klassenübergreifende Leistungsstandards herangezogen, so dass Vergleiche mehr oder weniger unabhängig vom jeweiligen Klassenleistungsniveau sind.

Der Einfluss des Klassenkontexts auf die Benotung kann als unabhängig von der Klassenstufe gesehen werden. Zwar liegen Befunde vor, nach denen die Notengebung in unteren Klassen insgesamt milder ausfällt als in höheren Stufen und die Noten – insbesondere die Rechtschreibnoten – vom 2. bis 6. Schuljahr deutlich abfallen (Thiel & Valtin, 2002; Trautwein & Baeriswyl, 2007¹), dennoch kann davon ausgegangen werden, dass diese „Verschiebung“ der Notenskala keinen Einfluss auf die Berücksichtigung des Klassenkontexts bei der Benotung hat. Die Noten fallen mit steigender Klassenstufe für alle Kinder schlechter aus.

Fragestellung

Die vorliegende Studie untersucht zu einem früheren Zeitpunkt als in den Arbeiten von Trautwein und Baeriswyl (2007) bzw. Tiedemann und Billmann-Mahecha (2007) den Einfluss des Klassenleistungsniveaus auf die Benotung. Die Ergebnisse von Trautwein und Baeriswyl (2007) beleuchten die Bewertungspraxis in 6. Klassen, die Studie von Tiedemann und Billmann-Mahecha (2007) bezieht sich auf Übergangsempfehlungen am Ende der 4. Klasse – in der vorliegenden Studie wird der Einfluss des Klassenleistungsniveaus auf die Benotung am Ende der 3. Klasse betrachtet.

Anders als in bisherigen Arbeiten, in denen verschiedene Leistungsbereiche zusammengefasst wurden, soll hier für die Bereiche Lesen und Rechtschreiben getrennt untersucht werden, in wieweit die Lese- und Rechtschreibnote vom Klassenleistungsniveau beeinflusst ist.

Nach den bisher berichteten Befunden ist zu erwarten, dass sowohl die individuelle Leistung als auch das Leistungsniveau der Klasse die Note beeinflusst. Auf individueller Ebene kann angenommen werden, dass mit steigender Testleistung die Bewertung durch die Lehrperson positiver ausfällt, während auf Klassenebene eine gute Testleistung in leistungsstarken Klassen mit schlechterer individueller Bewertung verknüpft ist als in leistungsschwachen Klassen.

Methode

Stichprobe

Die vorliegende Studie ist Teil des Projekts EVER (Entwicklung eines Vorschulscreenings zur Erfassung von Risikokindern für Sprach- und Schriftspracherwerbsprobleme)², in dessen Rahmen 1'256 Mannheimer Drittklässler/-innen am Ende des Schuljahres 2005/06 mit Hilfe von Lese- und Rechtschreibtests untersucht wurden. Von 1'243 Kindern liegen zudem Noten für Lesen und Rechtschreiben sowie die Gesamtnote im Fach Deutsch vor.

Von den insgesamt 90 Klassen konnten aufgrund zu geringer Anzahlen untersuchter Kinder in den jeweiligen Klassen für die folgenden Analysen nur 66 bzw. 67 Klassen berücksichtigt werden. In jeder der berücksichtigten Klassen liegen vollständige Daten von mindestens 10 Kindern vor³. Darüber ergeben sich die in Tabelle 1 dargestellten Stichprobengrößen. Die reduzierte Stichprobe kann als repräsentativ für Mannheim gelten (siehe Tabelle 1).

Tabelle 1: Stichprobe

	<i>N</i> Klassen	<i>N</i> Klassen	Durchschnittliche Anzahl berücksichtigter Kinder pro Klasse	Mädchenanteil (in %)	Deutsch als Muttersprache (in %)
Ursprüngliche Stichprobe	90	1'256	12.6	51.3	78.7
Rechtschreibung	66	1'071	16.2	51.8	82.5
Lesegeschwindigkeit	67	1'082	16.2	51.8	82.6
Leseverstehen	66	1'069	16.2	51.9	83.2

Abhängige Variablen

Als abhängige Variablen werden die Noten im Lesen und Rechtschreiben sowie die Gesamtnote im Fach Deutsch herangezogen. Die Deutschnote entspricht der Zeugnisnote am Ende der 3. Klasse. Bei der Bewertung der Lese- und Rechtschreibleistungen konnten die Lehrkräfte auch halbe Noten (Zwischennoten) vergeben.

Prädiktorvariablen

Als Prädiktorvariablen werden die mittels standardisierter Verfahren ermittelten Lese- und Rechtschreibleistungen verwendet.

Die Rechtschreibkenntnisse wurden mit Hilfe des Diagnostischen Rechtschreibtests für 3. Klassen (*DRT 3*, Müller 2004) erfasst, einem 44 Wörter umfassenden Lückendiktat.

Im Bereich Lesen wurden die Lesegeschwindigkeit, die Vorläuferfertigkeiten für das verstehende Lesen sowie das Leseverstehen erfasst. Durchgeführt wurden

- die Würzburger Leise Leseprobe (*WLLP*, Küspert & Schneider, 1998) zur Erfassung der Lesegeschwindigkeit sowie
- Knuspels Leseaufgaben (*Knuspel-L*, Marx, 1998) zur Erfassung der Vorläuferfertigkeiten für das verstehende Lesen (Score 1: Hörverstehen, Rekodieren, Dekodieren) und des Leseverstehens (Score 2: Rekodieren, Dekodieren, Leseverstehen).

Die individuelle Leistung wird als Prädiktor auf der Individualebene verwendet. Auf der Klassenebene fließt das über alle Schüler/-innen einer Klasse aggregierte Leistungsniveau der Klasse als Prädiktor ein. Verwendet werden jeweils die *T*-Werte. Die *T*-Werte der *WLLP* entsprechen nicht den *T*-Werten aus den Normtabellen – die dort für Mädchen und Jungen getrennt erstellten Normen sind für die vorliegende Fragestellung nicht sinnvoll zu verwenden. Daher wurden aus den ebenfalls im Testhandbuch angegebenen Statistiken der 3. Klasse *T*-Werte berechnet.

Statistische Analysen

Um der hierarchischen Datenstruktur gerecht zu werden, werden Mehrebenenanalysen gerechnet. Dazu wird das Softwarepaket HLM 6.04 (Raudenbush, Bryk & Congdon, 2007) verwendet. Die berichteten Modelle sind random-intercept-Modelle.

Ergebnisse

Schulnoten, Testleistungen und ihre Beziehungen

Die mittleren Noten für Rechtschreiben, Lesen und das Fach Deutsch sind in Tabelle 2 dargestellt. Wie bei Thiel und Valtin (2002) wird die Rechtschreibleistung am strengsten und die Leseleistung am mildesten beurteilt. In der vorliegenden Untersuchung fallen allerdings die Notenmittelwerte fast eine halbe Note schlechter als die Notenmittelwerte der 3. Klasse bei Thiel und Valtin (2002) aus – sie entsprechen ungefähr den von Thiel und Valtin (2002) ermittelten Notenmittelwerten der 4. Klasse bzw. 5. Klasse. Damit sind die hier vergebenen durchschnittlichen Noten mit der Lernstandseinschätzung bei Trautwein und Baeriswyl (2007) vergleichbar.

Am höchsten fallen die Zusammenhänge zwischen den Noten untereinander aus – der höchste Zusammenhang ergibt sich zwischen der Deutsch- und der Rechtschreibnote (siehe Tabelle 2). Zwischen den Lese- und Rechtschreibtestleistungen bestehen geringere Korrelationen. Der hohe Zusammenhang von $r = .93$ zwischen den beiden Scores des *Knuspel-L* erklärt sich über die in beiden Scores berücksichtigten Rekodier- und Dekodierleistungen.

Tabelle 2: Mittelwerte, Standardabweichungen und Produkt-Moment-Korrelationen

		Korrelationen									
		<i>M</i>	<i>SD</i>	<i>Min</i>	<i>Max</i>	(1)	(2)	(3)	(4)	(5)	(6)
Noten	(1) Rechtschreiben	2.82	1.25	1	6						
	(2) Lesen	2.44	.93	1	5	.73					
	(3) Deutsch	2.59	.97	1	5	.87	.79				
Testleistungen	(4) <i>DRT 3</i>	49.72	9.47	25.84	69.23	-.74	-.66	-.70			
	(5) <i>WLLP</i>	48.29	10.45	4.90	71.47	-.52	-.53	-.52	.54		
	(6) <i>Knuspel-L Score 1</i>	47.02	10.04	10.00	73.00	-.58	-.55	-.61	.65	.50	
	(7) <i>Knuspel-L Score 2</i>	48.02	9.52	18.00	72.00	-.62	-.59	-.64	.68	.52	.93

Alle Korrelationen sind signifikant bei $p < .001$.

Die Verteilungen der Testleistungen und Noten sind nicht identisch. Während die Testleistungen annähernd normal verteilt sind, folgen die Noten einer rechtsschiefen Verteilung. Da deswegen die maximal mögliche Korrelation nicht $r = 1$ beträgt, wurden die Korrelationen zwischen Testleistungen und Noten an der jeweils maximal möglichen Korrelation korrigiert. Es ergeben sich mittlere bis hohe Zusammenhänge (siehe Tabelle 3). Die Korrelation zwischen der Rechtschreibtestleistung und den Noten fällt am höchsten aus.

Tabelle 3: Korrigierte Korrelationen. Die nicht angegebenen Korrelationen wurden nicht korrigiert.

		Korrelationen					
		(1)	(2)	(3)	(4)	(5)	(6)
Noten	(1) Rechtschreiben						
	(2) Lesen						
	(3) Deutsch						
Testleistungen	(4) <i>DRT 3</i>	-.78	-.70	-.76			
	(5) <i>WLLP</i>	-.55	-.57	-.57			
	(6) <i>Knuspel-L Score 1</i>	-.61	-.59	-.66			
	(7) <i>Knuspel-L Score 2</i>	-.65	-.62	-.68			

Mehrebenenanalysen

Einfluss von individueller Rechtschreibleistung und Klassenleistung auf die Rechtschreibnote

Die Rechtschreibnote wird sowohl von der individuellen Rechtschreibleistung im *DRT 3* als auch vom Leistungsniveau der Klasse beeinflusst (siehe Tabelle 4). Der individuellen Leistung kommt dabei deutlich größeres Gewicht zu als dem Klassenleistungsniveau: Je besser die Leistung im *DRT 3*, desto besser fällt die Benotung durch die Lehrkraft aus. Der Einfluss des Klassenleistungsniveau fällt wie erwartet aus: Gleiche Testleistungen werden in leistungsstarken Klassen schlechter bewertet als in leistungsschwachen Klassen. Dieser Effekt ist mit $\beta = .28$ und einer Effektgröße von $\Delta = .36$ jedoch relativ gering und deutlich kleiner als in der 6. Klasse (Trautwein & Baeriswyl, 2007, dort beträgt $\Delta = -1.10^4$). Die Effektgröße von $\Delta = .36$ bedeutet, dass in zwei Klassen, die sich in ihrem Leistungsniveau um zwei Standardabweichungen unterscheiden, die gleiche Leistung in der leistungsschwächeren Klasse um .36 besser benotet wird als in der leistungsstarken Klasse. Die Varianzaufklärung liegt bei 58.6% und damit in einer ähnlichen Größenordnung wie die 67% Varianzaufklärung bei Trautwein und Baeriswyl (2007).

Tabelle 4: Standardisierte Regressionsgewichte für individuelle Leistung und Klassenleistung

		Rechtschreibnote	Lesenote	Deutschnote
		β	β	β
Rechtschreibung (<i>DRT 3</i>)	Indiv. Leistung	-.79	-	-.46
	Klassenleistung	.28	-	.36
Lesegeschwindigkeit (<i>WLLP</i>)	Indiv. Leistung	-	-.33	-.16
	Klassenleistung	-	n.s.	n.s.
Vorläuferfertigkeiten (<i>Knuspel-L Score 1</i>)	Indiv. Leistung	-	n.s.	n.s.
	Klassenleistung	-	n.s.	n.s.
Leseverstehen (<i>Knuspel-L Score 2</i>)	Indiv. Leistung	-	-.46	-.24
	Klassenleistung	-	n.s.	n.s.

Einfluss von individuellen Lesefähigkeiten und Klassenleistung auf die Lesenote

Die individuelle Lesegeschwindigkeit und das individuelle Leseverstehen schlagen sich in der Lesenote nieder. Je schneller gelesen wird und je besser ausgeprägt das Leseverstehen ist, desto besser fällt die Benotung der Leseleistung aus. Die Vorläuferfertigkeiten spielen bei der Bewertung der Leseleistung keine Rolle. Auch die Klassenleistungsniveaus aller Teilkomponenten wirken sich nicht auf die Benotung der Leseleistung aus. Mit Hilfe der berücksichtigten Variablen lässt sich 46.1% der Varianz aufklären.

Einfluss von individuellen Lese-Rechtschreibleistungen und Klassenleistung auf die Deutschnote

Die individuelle Rechtschreibleistung beeinflusst die Deutschnote positiv (siehe Tabelle 4). Je besser die Rechtschreibleistung im *DRT 3*, desto besser fällt die Deutschnote aus. Auch die Lesegeschwindigkeit und das Leseverstehen werden in der Deutschnote berücksichtigt. Die Vorläuferfertigkeiten spielen dagegen für die Deutschnote keine bedeutsame Rolle. Zudem wird die Deutschnote negativ beeinflusst vom Klassenleistungsniveau beim Rechtschreiben (siehe Abbildung 1). Leistungsgleiche Kinder erhalten in starken Klassen hinsichtlich der Rechtschreibleistung eine schlechtere Note als in Klassen, deren durchschnittliche Rechtschreibleistung geringer ausfällt. Mit einer Effektgröße von $\Delta = .53$ fällt allerdings auch dieser Effekt kleiner aus als bei den von Trautwein und Baeriswyl (2007) untersuchten Kindern in 6. Klassen. Die durchschnittlichen Lesefertigkeiten einer Klasse wirken sich auf die Deutschnote nicht aus. Mit Hilfe der berücksichtigten Variablen lässt sich 58.7% der Varianz aufklären.

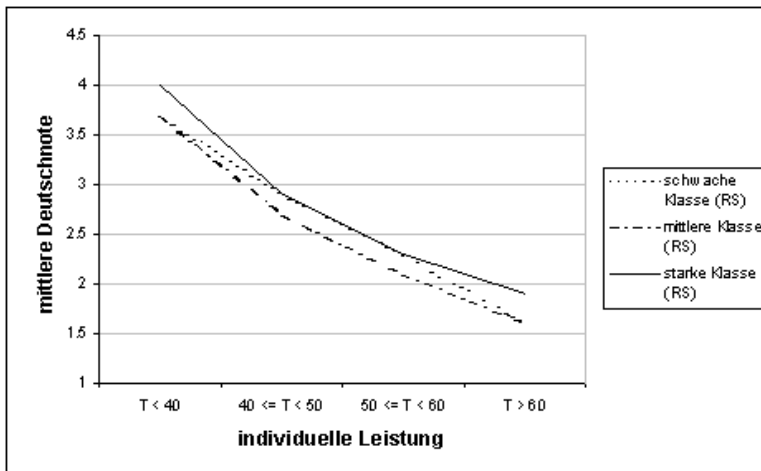


Abbildung 1: Mittlere Deutschnote bei gleicher individueller Leistung in rechtschreib-schwachen Klassen (mittlerer T-Wert < 45, N = 11), in mittelmäßigen Klassen (mittlerer T-Wert zwischen 45 und 55, N = 44) und rechtschreib-starken Klassen (mittlerer T-Wert > 55, N = 11)

Diskussion

In der vorliegenden Arbeit sollte untersucht werden, welchen Einfluss das Leistungsniveau einer Klasse auf die Benotung am Ende des 3. Schuljahres hat. Aufgrund der in anderen Studien nachgewiesenen Referenzgruppeneffekte bei der

Bewertung von Leistung war anzunehmen, dass die Schulnoten den in standardisierten Tests festgestellten Leistungsstand eines Kindes nur mangelhaft wiedergeben und von der Leistungsstärke der entsprechenden Klasse beeinflusst sind.

Die vorliegenden Ergebnisse zeichnen kein ganz so negatives Bild. Zwar zeigt sich, dass sowohl die Rechtschreib- wie auch die Deutschnote vom Leistungsniveau einer Klasse im Rechtschreiben beeinflusst ist, jedoch in weit geringerem Ausmaß als aufgrund vorheriger Studien angenommen werden musste. Nach den vorliegenden Befunden hat der Leistungsstand einer Klasse im Lesen keine Auswirkungen auf die Lesenote. Alle Noten sind vor allem vom individuellen Leistungsstand der Schüler/-innen beeinflusst. In der methodisch ähnlichen Arbeit von Trautwein und Baeriswyl (2007) beeinflusst das Klassenleistungsniveau die Bewertung am Ende der 6. Klasse in sehr viel stärkerem Maße als in der vorliegenden Arbeit.

Warum fällt der Einfluss des Klassenleistungsniveaus auf die Benotung im 6. Schuljahr deutlich höher aus als im 3. Schuljahr?

Eine mögliche Erklärung könnte in der jeweils unterschiedlichen Zeit der Beschulung zum Zeitpunkt der Studien und deren Auswirkung zu suchen sein. Die bei Trautwein und Baeriswyl (2007) untersuchten Schüler/-innen wurden sechs Jahre lang im mehr oder weniger gleichen Klassenkontext beschult, bevor sie vor der 7. Klasse auf unterschiedliche Schularten aufgeteilt wurden. Innerhalb der ersten sechs Jahre änderten sich demnach die Mitschüler/-innen kaum – jedes Kind blieb in einem ähnlichen Leistungsumfeld und hatte somit in Hinblick auf die Einschätzung der eigenen Fähigkeiten eine konstante Vergleichsbasis. Eine eher stabile Leistungshierarchie der Klasse könnte eine wenig flexible Selbst- und Fremdbewertung bedingen. Weder für Schüler/-innen noch für Lehrkräfte ergibt sich bei überwiegend stabilen Rangplätzen ein Anlass, getroffene Bewertungen in Frage zu stellen. Bei der Beurteilung am Ende des Schuljahres besteht dann die Gefahr, dass Lehrkräfte diese seit Jahren bestehende Hierarchie in die Notengebung mit einbeziehen, ohne mögliche Erweiterungen dieser Rangreihe außerhalb der Klasse zu berücksichtigen. In Klassen, die seit drei Jahren bestehen, sind diese Prozesse möglicherweise noch nicht so stark ausgeprägt.

Denkbar ist auch, dass Lehrkräfte aufgrund anderer Klassenkontextmerkmale zu einem Urteil über die Leistungsfähigkeit einer Klasse gelangen und dies in ihren Noten berücksichtigen. Wenn diese Klassenkontextmerkmale in tatsächlichem Zusammenhang mit dem Leistungsniveau einer Klasse stehen, kann die Benotung den individuellen Leistungsstand unabhängig vom Leistungsniveau der Klasse wiedergeben. So könnte in einer Klasse mit hohem Anteil an Kindern mit Migrationshintergrund oder mangelnden Deutschkenntnissen (unter der Annahme, dass diese Variablen das Klassenleistungsniveau insgesamt beeinflussen) die Lehrkraft aufgrund dieser Merkmale vom geringen Leistungsniveau ihrer Klasse überzeugt sein und es in den Noten berücksichtigen. Nicht auszu-

schließen ist auch, dass die Wirkung dieser Variablen erst über Erwartungseffekte entsteht.

Schließlich könnten auch die eingesetzten Verfahren für die unterschiedlichen Befunde verantwortlich sein. In der vorliegenden Studie wurde die übliche Notenskala verwendet, bei Trautwein und Baeriswyl (2007) kam dagegen eine vierstufige Skala zum Einsatz. Zudem wurden in der vorliegenden Studie die Bereiche Lesen und Rechtschreiben getrennt erfasst, während bei Trautwein und Baeriswyl (2007) Lernstandsbeurteilungen in den Bereichen Mathematik und Deutsch für die Analyse zusammengefasst wurde.

Weitere Forschung zur Erklärung der unterschiedlichen Befunde sowie zum Einfluss möglicher Klassenkontextmerkmale auf das Leistungsniveau einer Klasse erscheint nötig.

Zur Strenge der Beurteilung in der 3. Klasse

Thiel und Valtin (2002) berichten von einer strengeren Bewertung in höheren Klassen und einem besonders starken Einbruch der Noten am Ende der 4. Klasse. Dabei ist zu berücksichtigen, dass die Daten von Thiel und Valtin (2002) in Berlin erhoben wurden – dort findet der Übergang auf die weiterführende Schule erst nach der 6. Klasse statt. In Bundesländern wie Baden-Württemberg, wo der Übergang bereits nach der 4. Klasse vollzogen wird, ergibt sich zu einem früheren Zeitpunkt die Notwendigkeit, Eltern und Schüler/-innen über eine mögliche Bildungsempfehlung zu informieren. Da die in Baden-Württemberg bindende Bildungsempfehlung bereits zum Halbjahr der 4. Klasse ausgestellt wird, muss am Ende der 3. Klasse durch die Zeugnisnoten schon ein entsprechendes Signal gesetzt werden. In Berlin kann dies in späteren Schuljahren geschehen. Diese Annahme bestätigt sich in der vorliegenden Studie: Anders als bei Thiel und Valtin (2002) fällt die Benotung am Ende der 3. Klasse nicht auffallend mild aus. Die durchschnittlich vergebenen Noten entsprechen den bei Thiel und Valtin (2002) berichteten Notendurchschnitten am Ende der 4. Klasse bzw. 5. Klasse sowie den durchschnittlichen Leistungseinschätzungen am Ende der 6. Klasse bei Trautwein und Baeriswyl (2007). Festzustellen bleibt allerdings auch, dass das von Thiel und Valtin (2002) beschriebene Muster bestätigt werden kann: Die Leseleistung wird am mildesten bewertet, die Rechtschreibleistung am strengsten.

Dieses Ergebnis spricht dafür, dass nicht die Klassenstufe für die Strenge der Bewertung verantwortlich ist, sondern die zeitliche Nähe zum Übergang in die weiterführende Schule. Eine weitere Bestätigung für diese Hypothese kann darin gesehen werden, dass die Noten in der 2. Klasse der hier untersuchten Kinder noch bedeutsam besser ausfallen. Insbesondere die Deutsch- und die Rechtschreibleistung wird signifikant milder bewertet, so dass die mittleren Noten um knapp eine Viertel Note über den durchschnittlichen Noten der 3. Klasse liegen.

Zur Gewichtung von Rechtschreiben und Lesen in Deutschnoten

Die berichteten Ergebnisse verdeutlichen auch, welche Fertigkeiten die Lehrkräfte in ihren Noten berücksichtigen. Zur Bewertung der Leseleistung wird neben der Lesegeschwindigkeit insbesondere das Leseverstehen eines Kindes herangezogen. Dagegen spielen die Vorläuferfertigkeiten für das verstehende Lesen keine Rolle bei der Benotung. Dies ist insofern wenig erstaunlich, als die Vorläuferfertigkeiten hauptsächlich für den Leseerwerbsprozess von Bedeutung sind. Es ist davon auszugehen, dass der Großteil der Kinder am Ende der 3. Klasse einigermaßen sicher lesen kann, so dass den Vorläuferfertigkeiten keine diskriminierende Funktion mehr zukommt. Demgegenüber ist das Leseverstehen zunehmend von Bedeutung, so dass in der Lesenote berücksichtigt wird, in welchem Ausmaß diese neue Aufgabe gemeistert wird.

Auch für die Deutschnote spielen die Vorläuferfertigkeiten demnach erwartungsgemäß keine Rolle. Dafür fließen in die Deutschnote die jeweiligen Rechtschreibkenntnisse eines Kindes ein. Rechtschreibung scheint für die Lehrkräfte die zentrale Fähigkeit im Fach Deutsch zu sein. Sicherlich ist diese starke Gewichtung der Rechtschreibleistung bei der Bewertung der Leistung im Fach Deutsch auch damit begründet, dass diese Fähigkeit deutlich wahrnehmbar ist. So wird in den Diktaten und Aufsätzen die Rechtschreibfähigkeit offensichtlich und ist damit standardisierter und objektiver erfassbar als die Lesefähigkeit. Die Bewertung von Diktaten legt einen sozialen Vergleich nahe. Dies wird auch deutlich in den Varianzen von Rechtschreib- und Lesenoten: Die Rechtschreibnoten streuen stärker, das Notenspektrum wird also eher ausgeschöpft als bei der Bewertung der Lesefähigkeit.

Problematisch ist dieses Ergebnis vor dem Hintergrund, dass mangelnde Rechtschreibleistungen mit Hilfe gezielter und stetiger Übung eher ausgeglichen werden können als mangelnde Lesefähigkeit bzw. die dazu nötigen Verarbeitungsprozesse. Die eigentlich wichtigere Fähigkeit Lesen wird jedoch von den Lehrkräften offensichtlich als weniger wichtig eingeschätzt und in den Deutschnoten kaum berücksichtigt. Allerdings fehlen in Grundschulen vermutlich auch Methoden zur objektiven Erfassung der Leseleistung. Die Schwierigkeiten der Lehrkräfte, die Lesefähigkeit in ähnlich standardisierter Weise zu erfassen wie die Rechtschreibfähigkeit im Diktat, spiegeln sich in der Gewichtung der Leseleistung in der Deutschnote wieder.

Die Varianzaufklärung zeigt, dass die Noten nicht ausschließlich auf der individuellen Leistung und dem durchschnittlichen Abschneiden der besuchten Klasse beruhen. Es muss demnach angenommen werden, dass in die Noten weitere Informationen und Beobachtungen der Lehrkraft einfließen.

Zur Problematik von Noten

Die Tatsache, dass Lehrkräfte in die Noten weitere Informationen und Beobachtungen einfließen lassen, muss nicht negativ sein. So beruhen Noten auf schrift-

licher und mündlicher Leistung und einem längeren Zeitraum als die in einem einmalig durchgeführten Test erfasste Leistung, die möglicherweise von Tagesform, Aufregung und dergleichen beeinträchtigt ist (Schrader & Helmke, 2001). Zudem kann in den Schulnoten das Anspruchsniveau im Unterricht berücksichtigt werden. Es ist anzunehmen, dass ein Kind, das im Schulleistungstest schlecht abschneidet, jedoch auch in einer durchschnittlich schwachen Klasse unterrichtet wird, bei entsprechender Förderung eine bessere Leistung erzielen könnte. Diese mangelnde Förderung oder der geringe Anregungsgehalt des Unterrichts kann in Noten, in die der Klassenkontext in starkem Maße einfließt, eher berücksichtigt werden, so dass dann die Noten das Potenzial eines Kindes wiedergeben können. Problematisch ist dies allerdings für „schwache“ Kinder in leistungsstarken Klassen, denn in diesen Klassen unterschätzen Noten, die den Klassenkontext in starkem Maße berücksichtigen, die tatsächlich erbrachte Leistung.

In der Notenverordnung des Ministeriums für Kultus, Jugend und Sport Baden-Württemberg werden Noten anhand einer kriterialen Bezugsnorm definiert, ohne konkrete Standards zu nennen (Rheinberg, 2001). Fehlende Standards können demnach auch Vorteile haben, insbesondere in Fällen, in denen als Ursache für das schwache Abschneiden im Leistungstest die Qualität des Unterrichts herangezogen werden muss.

Hier stellt sich eine grundsätzliche Frage: In welchem Ausmaß sollte eine Lehrkraft das Leistungsniveau einer Klasse bei der Benotung berücksichtigen? Findet das Klassenleistungsniveau keine Berücksichtigung, orientiert sich die Lehrkraft bei der Benotung also voll und ganz an den Vorgaben im Curriculum und damit an einer kriterialen Bezugsnorm, besteht die Gefahr, dass zwar der tatsächliche Leistungsstand, nicht jedoch die eigentliche Leistungsfähigkeit benotet wird. Bei idealen Rahmenbedingungen, zu denen auch der Unterricht zählt, könnte ein Kind eventuell sehr viel bessere Leistung zeigen als unter weniger günstigen Bedingungen. Eine Berücksichtigung des Klassenleistungsniveaus bei der Benotung könnte demnach darauf schließen lassen, dass die Lehrkraft das Anspruchsniveau im eigenen Unterricht kritisch beleuchtet und dabei feststellt, dass ihr Unterricht und / oder die Rahmenbedingungen nicht ideal für die Leistungsentwicklung der Kinder sind. Fließt das Leistungsniveau einer Klasse in die Benotung ein, ergibt sich das Problem, dass die Noten den objektiven Leistungsstand eines Kindes nicht wiedergeben. Da eine soziale Bezugsnorm gewählt wird, die nicht wie bei der kriterialen Bezugsnorm über die einzelnen Klassen hinausgeht, sind die Noten über einzelne Klassen nicht vergleichbar.

Dieses Dilemma ist nicht lösbar. Der Idealfall, bei dem die Noten den individuellen Leistungsstand wiedergeben, ohne vom Leistungsniveau der Klasse beeinflusst zu sein, setzt voraus, dass die Bildungsbedingungen im Unterricht in allen Klassen vergleichbar (und möglichst gut) ausgeprägt sind. Davon kann in der Realität nicht ausgegangen werden. Selbst Parallelklassen einer Schule, bei denen angenommen werden kann, dass die Schüler-*innenschaft* vergleichbar ist, unter-

scheiden sich in ihrem Leistungsniveau – diese Unterschiede können nur von Merkmalen des Unterrichts bzw. der Lehrkraft herrühren. Vor diesem Hintergrund muss eine Bewertung der dargestellten Ergebnisse offen bleiben.

Anmerkungen

- 1 Der bei Trautwein und Baeriswyl (2007) berichtete Mittelwert aus den Lernstandseinschätzungen der Lehrkräfte am Ende der 6. Klasse liegt bei 2.47 und damit circa eine Viertelnote über dem bei Thiel und Valtin (2002) angegebenen Mittelwert der Deutsch- und Mathenoten am Ende des 6. Schuljahres. Dabei muss jedoch berücksichtigt werden, dass bei Trautwein und Baeriswyl (2007) nur eine vierstufige Skala zum Einsatz kam (mit Bezeichnungen, die der Notenskala entsprechen), während bei Thiel und Valtin (2002) die übliche sechsstufige Notenskala verwendet wurde. Mit dieser Einschränkung kann das Ergebnis von Trautwein und Baeriswyl (2007) als Bestätigung des Befunds von Thiel und Valtin (2002) gesehen werden.
- 2 Das Projekt EVER (Entwicklung eines Vorschulscreenings zur Erfassung von Risikokindern für Sprach- und Schriftspracherwerbsprobleme) wurde gefördert durch die Dürer-Stiftung Hamburg, die Günter Reimann-Dubbers-Stiftung Heidelberg und die Pädagogische Hochschule Heidelberg.
- 3 Voranalysen haben ergeben, dass sich die Effekte ab 10 berücksichtigten Kindern pro Klasse nicht mehr bedeutsam verändern.
- 4 Anders als bei Trautwein und Baeriswyl (2007) ist die Effektgröße hier positiv. Dies ist darin begründet, dass die Noten nicht umkodiert wurden. Ein Anstieg der Noten bedeutet demnach eine schlechtere Bewertung.

Literatur

- Ingenkamp, K. (1969). *Zur Problematik der Jahrgangsklasse: eine empirische Untersuchung von Karlbeinz Ingenkamp*. Weinheim: Beltz.
- Küspert, P. & Schneider, W. (1998). *Würzburger Leise Leseprobe (WLLP)*. Göttingen: Hogrefe.
- Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79, 280-295.
- Marx, H. (1998). *Knuspels Leseaufgaben (KNUSPEL-L)*. Göttingen: Hogrefe.
- Ministerium für Kultus, Jugend und Sport Baden-Württemberg. Verordnung des Kultusministeriums über die Notenbildung (Notenbildungsverordnung, NVO). Zugriff am 19.4.2008. Verfügbar unter http://www.vd-bw.de/webvdbw/rechtsdienst.nsf/weblink/6631-21_02.n
- Müller, R. (2004). *Diagnostischer Rechtschreibtest für 3. Klassen (DRT 3)*. Göttingen: Hogrefe.
- Raudenbush, S., Bryk, A. & Congdon, R. (2007). *HLM for Windows, Version 6.04*. Chicago, IL: Scientific Software International.
- Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In: F. E. Weinert (Hrsg.), *Leistungsmessung in Schulen* (S. 59-72). Weinheim: Beltz.
- Scheib, K., Schöler, H., Fehrenbach, C., Roos, J. & Zöller, I. (2005). *Lese- und Rechtschreibtestleistungen am Ende der 1. und 2. Klasse - Ein Vergleich zweier Jahrgänge sowie eine Prüfung von Einflussfaktoren*. Heidelberg: Pädagogische Hochschule, Erziehungs- und Sozialwissenschaftliche Fakultät.
- Schrader, F.-W. & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Hrsg.), *Leistungsmessung in Schulen* (S. 45-58). Weinheim: Beltz.
- Tent, L. (2001). Zensuren. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 805-811). Weinheim: Beltz.
- Thiel, O. & Valtin, R. (2002). Eine Zwei ist eine Drei ist eine Vier. Oder: Sind Zensuren aus

- verschiedenen Klassen vergleichbar? In R. Valtin (Hrsg.), *Was ist ein gutes Zeugnis? Noten und verbale Beurteilungen auf dem Prüfstand* (S. 67-76). Weinheim: Juventa.
- Tiedemann, J. & Billmann-Mahecha, E. (2007). Zum Einfluss von Migration und Schulklassezugehörigkeit auf die Übergangsempfehlung für die Sekundarstufe I. *Zeitschrift für Erziehungswissenschaft*, 10, 108-120.
- Trautwein, U. & Baeriswyl, F. (2007). Wenn leistungsstarke Klassenkameraden ein Nachteil sind. Referenzgruppeneffekte bei Übertrittsentscheidungen. *Zeitschrift für Pädagogische Psychologie*, 21, 119-133.
- Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O. & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth grade mathematics. *Journal of Educational Psychology*, 98, 438-456.

Schlagnote: Big-Fish-Little-Pond-Effekt, individuelles Leistungsniveau, Klassenleistungsniveau, Benotung

L'influence du niveau de performance d'une classe sur la notation à la fin de la 3e année scolaire

Resumé

Les classes se distinguent les unes des autres par leurs niveaux de performances. Ce contexte différent a probablement des conséquences pour les élèves: non seulement leur image de soi est influencée par le niveau du groupe de référence («Big-Fish-Little-Pond-effect»), mais aussi la perception des élèves qu'ont les enseignants pourrait infléchir leur notation.

L'influence du groupe de référence sur la notation à la fin de la 3e année scolaire a été examinée pour 1'256 enfants. À partir du niveau de performance de la classe, les résultats des élèves à des tests standardisés de lecture et d'orthographe ont été mis en relation avec la notation des enseignants. L'analyse des données ne montre pas l'effet attendu du groupe de référence : le niveau de performance de chaque élève pèse davantage sur la notation que le niveau de performance général de la classe. La présente contribution questionne les raisons de ce résultat ainsi que la manière de considérer les effets du niveau de performance d'une classe dans le domaine de la notation.

Mots clés: Big-Fish-Little-Pond-effect, niveau de performance individuel, évaluation scolaire, jugement des enseignants, notation

L'influenza del livello di prestazione di una classe sulla valutazione alla fine del terzo anno scolastico.

Riassunto

Le classi si distinguono per il loro livello di prestazione, il che incide sugli allievi. Il gruppo di riferimento può influenzare non solo il concetto di sé degli allievi, ma anche il giudizio degli insegnanti (« Big-Fish-Little-Pond »).

L'influenza del gruppo di riferimento sulla valutazione alla fine del terzo anno di scuola è stata analizzata su 1256 bambini. Considerando il livello di prestazione della classe, i risultati ai test di lettura e di scrittura sono stati messi in relazione con la valutazione degli insegnanti. L'analisi dei dati non rivela l'effetto « Big-Fish-Little-Pond » atteso. Il livello individuale ha una influenza maggiore sulla valutazione che non il livello della classe.

Ciò porta a discutere le ragioni possibili di questo risultato e la questione fino a che punto il livello di prestazione di una classe debba essere preso in considerazione in ambito di valutazione.

Parole chiave: livello di prestazione individuale, valutazione scolastica, note scolastiche, giudizio degli insegnanti, Big-Fish-Little-Pond-effect

Influence of class achievement level on the grades at the end of the third grade

Abstract

School classes differ in their achievement level, with different contexts having varying effects on students. Not only is self concept influenced by the reference group (Big-Fish-Little-Pond-Effect), but teachers' grading can also be subject to reference group effects.

In this study, the influence of reference groups on grading at the end of third grade was investigated in a group of 1'256 children. The children's achievement in reading and spelling tests was related to their teachers' grading, taking the overall class achievement level into account.

Results do not show the expected reference group effect, but rather that individual achievement level influenced grading more than the overall class achievement level. Possible causes for this result, and the basic question as to what extent a group's overall achievement level should be taken into account in grading will be discussed.

Keywords: Class achievement, certification, Big-Fish-Little-Pond-Effekt, teacher grading