Empirische Sonderpädagogik, 2010, Nr. 1, S. 64-77

Lernverlaufsdiagnostik – Dargestellt am Beispiel der Entwicklung der Rechtschreibkompetenz in der Grundschule

Alfons Strathmann[1], Karl Josef Klauer[2], Michaela Greisbach[1]

[1]Universität zu Köln, [2]RWTH Aachen

In diesem Beitrag wird das Verfahren der Lernverlaufsdiagnostik im Bereich der Rechtschreibung weiter entwickelt und erstmalig getestet, wie sich die Erzeugung von Aufgabenstichproben durch einen Zufallsgenerator bewährt. Das Verfahren gestattet, den Lernverlauf einzelner Kinder wie ganzer Klassen über längere Zeiträume hinweg zu dokumentieren. Das geschieht am Beispiel der Beherrschung des Grundwortschatzes in sechs Grundschulklassen, wo ein halbes Schuljahr lang wöchentlich jeweils eine neue Zufallsstichprobe von zwanzig Wörtern generiert und als Diktat gegeben wurde. Tests dieser Art lassen sich nicht nach der klassischen Testtheorie analysieren, wohl aber nach dem Binomialmodell. Zusätzlich zu den typischen Lernverläufen bietet das Verfahren ein neues Maß für den Lernzuwachs sowie die Möglichkeit begründeter Prognosen des zukünftigen Lernerfolgs. Es resultieren zum Teil unerwartete Ergebnisse, so etwa, dass rund ein Drittel der Kinder in dem Halbjahr keinerlei Lernfortschritt an den Tag legte. Die Lernverlaufsdiagnostik kann Lehrkräften helfen, solche Entwicklungen eher zu erkennen. Allerdings sind weitere Forschungs- und Entwicklungsarbeiten zur Verbesserung der Technik erforderlich.

Schlüsselwörter: Lernverlaufsdiagnostik, Rechtschreibkompetenz, Grundschule

Curriculum Based Measurement – Demonstrated through the Development of Writing Competence in Primary School

In this contribution the curriculum based measurement is further developed and in the recent version empirically applied for the first time. The spelling competencies of children were regularly tested for a half year using a computer generated sample of twenty words belonging to a defined set of basic words. With such a procedure the classical test theory is not applicable but a binomial test model. The study proposes a new measure to assess the gains of learning and to predict future learning. Some of the results were not expected, particularly that nearly a third of the children did not improve at all during that period of time. This kind of diagnostic can foster the diagnostic compentencies of teachers. However, it is concluded that further appropriate research and development should be done.

Key words: curriculum based measurement, writing competences, primary school

Systematisch entwickelt wurde die Lernfortschrittsmessung von Deno in Minneapolis und vorwiegend im sonderpädagogischen Kontext eingesetzt (Deno, 1985; 2003; vgl. auch Fuchs, 2004). Dort läuft das Unternehmen unter der Bezeichnung „curriculum-based measurement" (abgekürzt CBM), um damit auszudrücken, dass es bei den zu testenden Inhalten um solche geht, die in der *jeweiligen Klasse* tatsächlich unterrichtet worden sind. Instrumente dieser Art sind also – im Gegensatz zu standardisierten Schulleistungstests – unmittelbar auf die in der fraglichen Klasse behandelten Inhalte bezogen. Ob man bei uns analog von curriculumbasierter Leistungsmessung sprechen sollte, mag unterschiedlich beurteilt werden. Walter (2009a) verwendet beispielsweise auch den Begriff der lernprozessbegleitenden Diagnostik.

Eine ausführlichere Darstellung der amerikanischen curriculumbasierten Messung einschließlich ihrer historischen wie bildungspolitischen Hintergründe findet man bei Klauer (2006), ferner bei Diehl und Hartke (2007) sowie bei Walter (2008; 2009a; 2009b). Walter hat überdies eine umfangreiche Datenerhebung zum sinnverständigen Lesen vorgelegt, einer Thematik, die im Arbeitskreis von Deno besonders intensiv erforscht worden ist. Dabei konnte Walter empirisch zeigen, dass Tests dieser Art wichtigen Gütekriterien entsprechen. Dieser Thematik wurde auch in den USA mit Recht umfangreiche Forschung gewidmet (vgl. etwa Van der Heyden, Witt, Naquin & Noell, 2001).

Ebenso griffen Strathmann und Klauer (2008) das Verfahren auf, den Leistungsstand der Lernenden im Hinblick auf ihre Rechtschreibleistungen wiederholt und möglichst längerfristig zu erheben. Sie entwickelten das Konzept insofern jedoch weiter, als sie zwei zusätzliche Bedingungen forderten, nämlich die *Kontent-* oder *Lehrzielvalidität* sowie das *Itemsampling*. Die erste Bedingung gewährleistet eine präzise Definition des jeweiligen Lehrziels durch die Definition der Menge von Aufgaben, welche die Lernenden später beherrschen sollen. Die Bedingung des Itemsamplings gewährleistet, immer neue Zufallsstichproben von Aufgaben aus der Grundmenge zu ziehen, um sicher zu sein, dass einerseits stets die Leistung bei ein und demselben Lehrziel gemessen wird, andererseits aber keine Testwiederholung stattfindet.

So konstruierte Testverfahren lassen sich nicht nach der klassischen Testtheorie behandeln. Retests gibt es nicht, und insofern lässt sich auch nicht die Reliabilität durch Testwiederholung ermitteln. Stattdessen werden aber immer wieder neue Aufgabenstichproben geboten und es ist üblich geworden, deren Ergebnisse zu interkorrelieren. Man spricht dann von der Paralleltestreliabilität, obgleich es sich nicht um Paralleltests im strengen Sinne handelt.

Auf Itemebene lassen sich weder Itemschwierigkeiten noch die Trennschärfen berechnen, weil es immer neue Items sind, die angeboten werden. Insofern kommt auch die Schätzung der inneren Konsistenz nicht in Frage, etwa Cronbachs α. Jedoch gilt für solche Verfahren das Binomialmodell, unter bestimmten zusätzlichen Bedingungen das Verallgemeinerte Binomialmodell oder das Betabinomialmodell (dazu ausführlich Klauer, 1987). So hatten schon Lord und Novick (1968, S. 234 ff; S. 523) am Beispiel gerade auch der Rechtschreibkompetenz gezeigt, dass ein Test, bei dem jeder Proband eine eigene zufällige Aufgabenstichprobe erhält, gemäß dem Binomialmodell auszuwerten ist und dass die Kuder-Ri-

charson-Formel 21 hierfür das geeignete Reliabilitätsmaß bietet. Bei einem Binomialtest stellt der Anteil richtiger Lösungen eine biasfreie Schätzung des Personenparameters dar. Handelt es sich um Daten einer ganzen Klasse, so fassen wir die Klasse als Individuum auf, das stets eine neue Itemstichprobe erhalten hat.

Im Folgenden wird erstmalig ein Zufallsgenerator eingesetzt, um bei jeder Testung eine neue Stichprobe von Aufgaben zu erzeugen. Dabei soll geprüft werden, welche Möglichkeiten sich mit dieser Vorgehensweise eröffnen, aber auch, welche Schwierigkeiten damit verknüpft sind. Das Verfahren dürfte auch für die pädagogische Praxis neue Wege eröffnen, etwa wenn Lehrkräfte Zugriff zu dem Aufgabengenerator haben und die Lernverlaufsdiagnostik in ihrer Klasse einsetzen können.

Fragestellung

Strathmann und Klauer (2008) hatten das Verfahren zur Diagnostik des Lernverlaufs im Bereich Rechtschreibung in drei Grundschulen und zwei Sonderschulen einer ersten empirischen Erprobung unterzogen, wobei noch Unzulänglichkeiten in Kauf genommen werden mussten: Die Grundmenge der Aufgaben umfasste nur 480 Wörter und die Kinder erhielten zwar immer neue Aufgabenstichproben, die allerdings noch nicht durch ein echtes Zufallsverfahren erzeugt werden konnten. Vielmehr wurden die zu diktierenden Wörter jeweils von studentischen Hilfskräften aus der Grundmenge ausgewählt. Der Einsatz des Verfahrens war auch zeitlich stärker begrenzt. Im folgenden Beitrag soll das nun weiter entwickelte Verfahren eingesetzt und erprobt werden: Insbesondere wurde die Aufgabenmenge erheblich erweitert, wie im Einzelnen unten beschrieben wird. Ferner wurde ein Zufallsgenerator eingesetzt, um jedes Mal eine neue zufällig gezogene Itemstichprobe zu erzeugen. Und schließlich wurde die Prozedur in den 2., 3. und 4. Klassen zweier Grundschulen ein halbes Schuljahr lang jede Woche eingesetzt, um Lernverläufe aufzeigen zu können.

Was die *Testqualität* betrifft, so hatten Klauer und Dänecke (1981) sowie Klauer (1984) an Beispielen zeigen können, dass durch Itemsampling erzeugte Testvarianten Paralleltests im klassischen Sinne sein können, sich also durch gleiche Mittelwerte, gleiche Varianzen und gleiche Kovarianzen auszeichnen. Im Rahmen der Verlaufsdiagnostik interessiert dabei besonders, ob die einzelnen Tests – hier die Diktate – einigermaßen gleich schwer sind. Wie bereits angedeutet, ist es nicht möglich, die Schwierigkeit der Aufgabenstichproben im Rahmen der Verlaufsdiagnostik direkt zu ermitteln, eben weil immer neue und zufällig gezogene Stichproben von Aufgaben eingesetzt werden. Hinzu kommt der Umstand, dass im Laufe der Zeit auch ein *Lernfortschritt* im Bereich der Rechtschreibung zu erwarten ist, der die Schwierigkeit einer einzelnen Itemstichprobe vermindern sollte, ein Umstand, dem bei der Schätzung der Testschwierigkeit Rechnung zu tragen ist.

Allerdings gilt es, noch einen weiteren Aspekt zu beachten. Lernverlaufsdiagnostik hat unvermeidlich mit *Veränderungsmessung* zu tun, will man doch den Lernzuwachs erfassen, mitunter sogar auch das Auf und Ab, das sich in manchen Fällen darstellen dürfte (Collins & Horn, 1991), wozu klassisch konstruierte Tests ohnedies kaum günstige Voraussetzungen bieten (Tack, 1986; Tent & Stelzl, 1993, S. 169-184; Petermann, 2001). Im vorliegenden Fall der längerfristigen Ver-

änderungsmessung ist daher zweierlei vorauszusetzen, gleiche Schwierigkeit der einzelnen Tests sowie änderungssensible Tests.

Die *Reliabilität* soll auf zweierlei Weise geschätzt werden. Die Paralleltestreliabilität wird ermittelt durch die Interkorrelationen der 20 Testdiktate. Da es sich hier jedoch um Binomialtests handelt, ist deren Reliabilität außerdem durch die Kuder-Richardson-Formel 21 (KR-21) zu berechnen (vgl. Klauer, 1987, S. 151f; de Gruijter & van der Kamp, 1984, S. 60; Lord & Novick, 1968, S. 523).

Hypothesen

Folgende Hypothesen sollen getestet werden.

H1: Für die mittlere Reliabilität M_R der Tests gilt $M_R \geq .75$. Begründung: Der Erwartungswert für die mittlere Reliabilität orientiert sich an den Erfahrungswerten, die in den USA und von Walter beim lauten Lesen erzielt worden sind und in aller Regel über dem Wert von $r_{tt} = 0.75$ liegen (vgl. Walter 2008; 2009a; 2009b).

H2: Mit den fortlaufend geschriebenen Testdiktaten steigt die mittlere Rechtschreibleistung an. Begründung: Mit dieser Hypothese wird der Lernfortschritt der Klasse, aber auch die Änderungssensibilität des Verfahrens erfasst. Daher wird getestet, ob die Regressionskoeffizienten in den sechs Klassen größer als null sind.

H3: Höhere Klassen beginnen auf einem höheren Leistungsniveau. Begründung: Mit dieser wie mit Hypothese 2 wird ein Aspekt der Validität des Verfahrens erfasst.

H4: Die unmittelbar aufeinander folgenden Testdiktate unterscheiden sich nicht in ihrer Schwierigkeit. Begründung: Für die Verlaufsdiagnostik ist es wichtig zu klären, ob die einzelnen Paralleltests einigermaßen gleich schwer sind, denn ungleich hohe Anforderungen können Änderungen im Lernverlauf vortäuschen, die es in Wirklichkeit so gar nicht gibt. Weil aber im Laufe der Zeit mit einem Lernzuwachs zu rechnen ist, werden nur die *unmittelbar* aufeinander folgenden Testdiktate auf Mittelwertsunterschiede verglichen, denn dabei sollte sich nur ein geringer Lernzuwachs einstellen.

Methode

Versuchspersonen

In die fortlaufenden Erhebungen waren 128 Kinder aus zwei Grundschulen mit zusammen sechs Klassen einbezogen, jeweils zwei Klassen des zweiten, dritten und vierten Schuljahrs. Kinder, die an drei oder mehr Terminen fehlten, wurden nicht berücksichtigt, so dass die Stichprobe letztlich aus 121 Kindern besteht.

Durchführung

In jeder Klasse wurde typischerweise jede Woche ein Rechtschreibtest erhoben, und zwar für 20 Wochen, also praktisch ein halbes Schuljahr lang von Februar bis zu den Sommerferien. Nur während der Osterferien entfielen die regelmäßigen Diktate. Die Erhebungen wurden von Studierenden des Lehramts für Sonderpädagogik durchgeführt, die sich bei der Durchführung genau an die Instruktionen hielten, damit vergleichbare Resultate erzielt werden konnten.

Die Diktate fanden in folgender Weise statt. Es waren immer nur einzelne Wörter zu schreiben, und zwar genau 20 Wörter pro Diktat. Die Wörter wurden nach fol-

gendem Muster diktiert. Zuerst wurde das Wort ausgesprochen, dann in einem ganzen Satz verwendet, um das Wortverständnis zu sichern, und schließlich wurden die Kinder aufgefordert, das Wort hinzuschreiben. Beispiel:
– *Wir schreiben jetzt das Wort „Vogel"*
– *Ein Vogel kommt geflogen.*
– *Schreibt jetzt: „Vogel".*

Das Verfahren wurde bereits von Klauer (1968) in Sonderschulen eingesetzt, ähnlich aber auch in den USA von Fuchs und Fuchs (1993). Es entlastet die Kinder sehr stark, fordert von ihnen nichts Zusätzliches außer der Leistung, um die es geht, nimmt wenig Zeit in Anspruch und entlastet die Auswertung von dem Problem, was mit Fehlern zu geschehen habe, die für die Fragestellung irrelevant sind. Bei der Auswertung wurde also nur festgestellt, ob die Schreibung richtig oder falsch war (vgl. Fuchs & Fuchs, 1993).

Material

Definition des Grundwortschatzes. Um lehrziel- oder kontentvalide Diktate zu ermöglichen, war es zunächst erforderlich, einen Grundwortschatz zu definieren, aus dem die zu diktierenden Wörter zu entnehmen waren. In einer Arbeitsgruppe von Studierenden unter der Leitung der Koautorin war in einem dreischrittigen Verfahren ein Grundwortschatz erstellt worden. Im ersten Schritt wurden drei publizierte Grundwortschatz-Listen gesichtet, nämlich die von Augst (1989), Naumann (1999) und Plickat (1987), wobei im Fall von Augst nur die Wörter des vierten Schuljahrs herangezogen wurden. Im zweiten Schritt wurde daraus ein Ausgangswortschatz von 2407 Wörtern gezogen. Aus diesem Bestand wurde im dritten Schritt schließlich ein Kernbestand ausgewählt, der aus der Schnittmenge der drei ursprünglichen Listen bestand, also die Wörter umfasst, die in allen drei Listen vertreten sind. Dieser Satz von 480 Wörtern galt fortan als unser Grundwortschatz. Es handelt sich dabei um Wörter aller Wortarten, die beispielsweise in Lesetexten für Kinder dieser Altersstufe besonders häufig auftreten und von den meisten Kindern ohne Schwierigkeiten verstanden werden. Die Wörter des Grundwortschatzes befinden sich alle in der Grundform (z. B. Infinitive bei Verben, Nominativ Singular bei Substantiven, Positiv bei Adjektiven). In einem zweiten Schritt wurden diese ergänzt um flektierte Formen (Substantive um den Nominativ Plural; Verben um 2. und 3. Person Singular Präsenz sowie um 3. Person Singular und Plural Präteritum und um das Partizip Perfekt; Adjektive um Komparativ und Superlativ). Somit erweiterte sich die Wortmenge, aus der die Stichproben gezogen wurden, auf 1334 Wörter.

Viele der Wörter haben in irgendeiner Form miteinander zu tun. Bei flektierten Verben ist das offensichtlich, etwa bei „gehen" und „gingen". In anderen Fällen existieren formale Ähnlichkeitsbeziehungen wie etwa in „Haus" und „aus", „lachten" und „brachten" usw. Insofern werden die zu erlernenden Wörter nicht gleiche Schwierigkeiten bieten. Aber das ist im Binomialmodell auch nicht gefordert, sofern Itemsampling stattfindet. Vom Lehrziel her soll natürlich die korrekte Schreibung all dieser Wörter beherrscht werden.

Erzeugung der Testmengen durch Itemsampling. Mittels eines Zufallsgenerators konnte für jedes Diktat eine neue Zufallsstichprobe von 20 Wörtern ohne Zurücklegen gezogen werden. Im Anhang befin-

det sich ein Beispiel für eine der Wortlisten.

Auswertung

Die Studierenden sammelten die Diktate der Kinder ein und zählten lediglich aus, wie viele der Wörter jeweils richtig geschrieben waren. Dabei wurde weder nach der Art noch nach der Anzahl der Fehlschreibungen pro Wort differenziert. Ein Wort war entweder richtig geschrieben oder nicht und nur die Richtigschreibungen wurden gezählt.

Anschließend trugen die Studierenden die Daten pro Test und Kind in ein Computerprogramm ein, das zugleich Lernverlaufskurven für jedes Kind erzeugte sowie eine Exceltabelle mit den Gesamtergebnissen der Klasse. Aufgrund der Gesamtergebnisse konnten dann auch Verlaufskurven für ganze Klassen erstellt werden.

Ergebnisse

Testung der Hypothesen

Hypothese 1 bezieht sich auf die Reliabilität des Verfahrens. Die Reliabilität soll – wie oben dargelegt - auf zweierlei Weise geschätzt werden. Die KR-21 hat den Vorteil, sich nur auf Daten des jeweiligen Tests zu beziehen, während die Paralleltestreliabilität hier als Mittelwert der 190 Korrelationen zwischen den 20 verschiedenen Tests erscheint und folglich nicht einen Test als solchen kennzeichnet. Aber da die Testdiktate durch Itemsampling erzeugt worden sind, sollten alle die gleiche Leistung messen.

Wie Tabelle 1 ausweist, kann Hypothese 1 nicht bestätigt werden. Nur in zwei der sechs Klassen werden mittlere Reliabilitäten in der erwarteten Größenordnung erzielt und betrachtet man die Mittelwerte über beide Schulen und alle Klassen hinweg, so ist der Befund enttäuschend. Tabelle 2 bietet die obere Hälfte der Korrelationsmatrix und damit einen Einblick in die tatsächlich erzielten Paralleltestreliabilitäten einer Klasse. Betrachtet man allerdings nur die Koeffizienten in Tabelle 2, die unmittelbar aufeinander fol-

Tab. 1: Reliabilitäten der Tests in den sechs Klassen

Schule	**Klasse**	**N**	**KR-21**	**Paralleltestreliabilität**
M	2	18	.76 ± .09	.77 ± .13
	3	22	.52 ± .20	.67 ± .19
	4	*25*	*.69 ± .10*	*.54 ± .21*
Mittelwert			.66	.66
B	2	12	.57 ± .24	.58 ± .24
	3	25	.77 ± .06	.80 ± .14
	4	22	.50 ± .35	.44 ± .38
Mittelwert			.61	.60

Tab. 2: Obere Hälfte der Korrelationsmatrix Grundschule Klasse 3 (D = Diktat)

	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
D1	.90	.66	.74	.71	.79	.51	.81	.80	.68	.48	.81	.48	.73	.73	.58	.88	.81	.78	.61
D2		.75	.73	.60	.69	.48	.71	.82	.63	.40	.67	.43	.74	.64	.50	.75	.58	.63	.47
D3			.85	.88	.79	.85	.73	.90	.78	.79	.71	.71	.70	.84	.78	.75	.36	.74	.68
D4				.86	.71	.78	.80	.93	.59	.53	.90	.68	.75	.87	.66	.83	.51	.90	.65
D5					.90	.91	.87	.87	.85	.85	.88	.85	.76	.96	.89	.86	.62	.91	.87
D6						.87	.90	.86	.97	.79	.85	.86	.74	.94	.93	.88	.71	.80	.88
D7							.76	.86	.83	.76	.80	.91	.65	.94	.90	.78	.41	.78	.84
D8								.85	.85	.70	.93	.84	.78	.92	.88	.80	.70	.85	.75
D9									.77	.62	.89	.76	.75	.92	.80	.88	.53	.83	.71
D10										.83	.73	.86	.72	.88	.93	.77	.63	.70	.85
D11											.55	.68	.45	.74	.88	.55	.35	.58	.66
D12												.82	.80	.94	.78	.91	.74	.95	.80
D13													.76	.92	.90	.71	.53	.78	.88
D14														.81	.59	.79	.77	.87	.82
D15															.91	.90	.66	.92	.90
D16																.71	.47	.70	.78
D17																	.81	.91	.86
D18																		.79	.77
D19																			.87

gende Diktate erzielt hatten, so ändert sich das Bild: Die 19 Koeffizienten der Hauptdiagonale von Tabelle 2 zeigen einen Mittelwert M = 0.81 ± 0.08 und einen Streubereich von 0.55 bis 0.91. In den anderen Klassen resultierten vergleichbare Werte. Offenbar variieren die Leistungen der Kinder *über die Zeit* stärker, als man dies erwarten mag.

Die Hypothesen 2 und 3 beziehen sich auf Aspekte der Validität des Verfahrens. Die Abbildungen 1 und 2 bieten einen Eindruck von den vorgefundenen Verhältnissen.

Gemäß Hypothese 2 sollten nur positive Regressionskoeffizienten zwischen der unabhängigen Variablen (den Testterminen) und der abhängigen Variablen (der mittleren Leistung) auftreten. Wie Tabelle 3 zeigt, ist das in der Tat der Fall. Von den sechs Regressionskoeffizienten sind fünf auch signifikant größer als null. Am schwächsten sind die Koeffizienten in den 4. Klassen, bei denen ohnedies ein Deckeneffekt zu erwarten war.

Nach Hypothese 3 ist zu erwarten, dass höhere Klassen einer Schule beim ersten Testdiktat schon auf einem höheren Leistungsniveau als niedere Klassen starten. Zur Prüfung dieser Hypothese wurde in jeder der Schulen eine Varianzanalyse gerechnet mit den Ergebnissen von Diktat 1 als der abhängigen und den drei Klassen als der unabhängigen Variablen. Beide Analysen fielen signifikant aus. In Schule M resultierte ein F (2; 63) = 42,85 (p < .01). Die nachfolgenden Kontraste belegten den Unterschied zwischen Klasse 2 und 3 (p < 0.01) und den Klassen 3 und 4 (p = .02). In Schule B brachte die ANOVA mit F (2; 57) = 584,1, p < .01, ein vergleichbares Ergebnis und hier unterschieden sich jeweils die höhere von der niederen Klasse ebenfalls bedeutsam (in beiden Fällen p < .01). Somit kann auch diese Hypothese weiterhin beibehalten werden.

Hypothese 4 erwartet homogene Testschwierigkeiten für die aufeinander folgenden Diktate. Da es sich jeweils um die gleichen Probanden handelt, werden t-Tests für abhängige Stichproben berechnet. Pro Klasse sind dann 19 Tests auf Mittelwertsunterschiede durchzuführen. Bei 19 Tests gegen das Signifikanzniveau von α = 0.05 sind allerdings auch bis zu drei

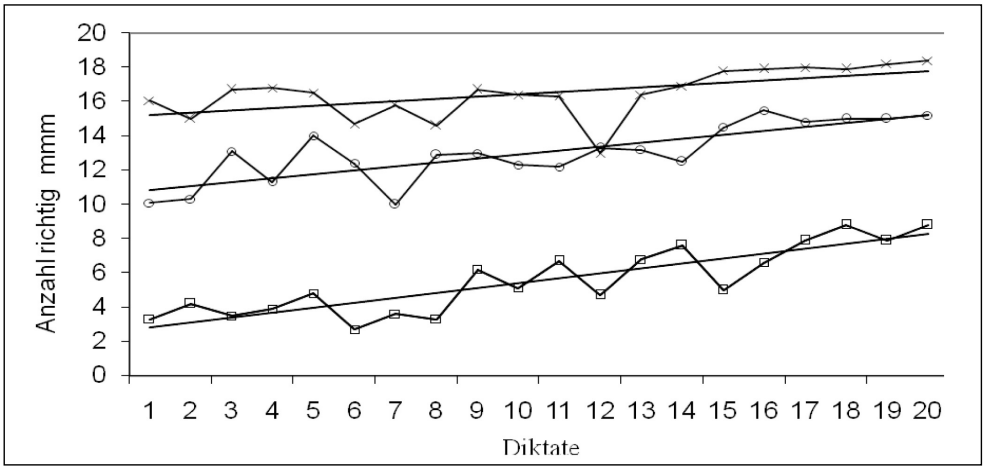

Abb. 1: Anstieg der durchschnittlichen Leistungen über ein halbes Schuljahr (untere Linie 2. Klasse, mittlere Linie 3. Klasse, obere Linie 4. Klasse) in Schule B

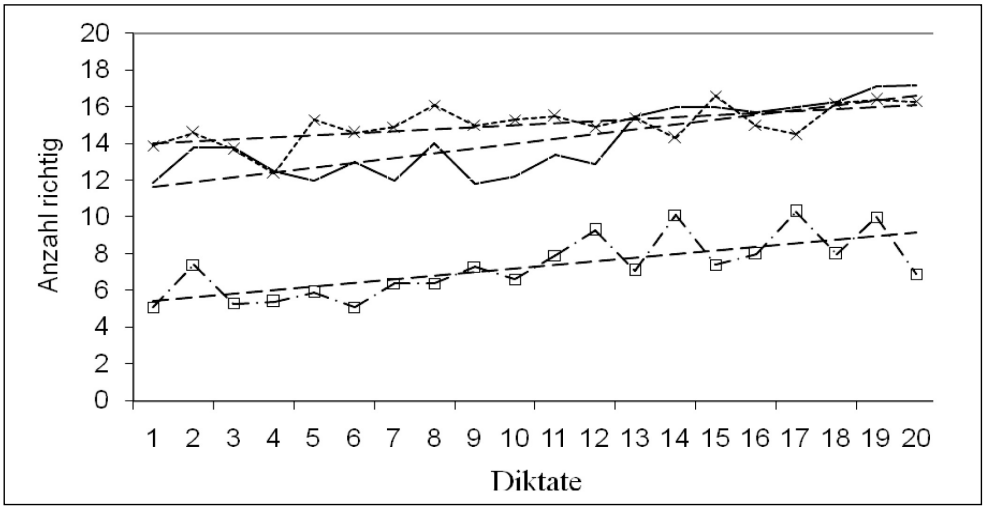

Abb. 2: Anstieg der durchschnittlichen Leistungen über ein halbes Schuljahr (untere Linie 2. Klasse, mittlere Linie 3. Klasse, obere Linie 4. Klasse) in Schule M

Zufallssignifikanzen zu erwarten (vgl. Tabelle S. 161 in Klauer, 2005). Daher sei zuvor die Bonferronikorrektur in der sequentiellen Variante von Holm (1979) vorgenommen. Man muss dabei allerdings auch beachten, dass dann eher β-Fehler unterlaufen können, also echte Mittelwertsunterschiede übersehen werden. Daher erhöhen wir das Signifikanzniveau auf $\alpha = 0.10$ und vergleichen den kleinsten p-Wert bei den t-Tests mit dem Wert 0.10 : 19 = 0.0053. Die Ergebnisse sind in Tabelle 4 dargestellt.

Die Größenordnung der Mittelwertsunterschiede lässt sich in der Variabilität um die Regressionslinien der Abbildun-

Tab. 3: Regressionskoeffizienten und Ergebnisse der Signifikanzprüfung

Schule	Klasse	Koeffizient	p
M	2	.25	<.01
	3	.26	<.01
	4	.11	<.01
B	2	.28	<.01
	3	.23	<.01
	4	.14	n. s.

Tab. 4: Anzahl signifikanter Mittelwertsdifferenzen

Schule	Klasse	N	Anzahl signifikanter Differenzen	Anzahl großer Effekte *d*
M	2	18	4	0
	3	22	6	2
	4	25	2	1
B	2	12	2	3
	3	25	6	1
	4	22	3	3

Anmerkung. Als groß gilt ein Effekt, wenn $d \mid \geq .80 \mid$

gen 1 und 2 grob einschätzen. Präzisere Information bietet das Maß der Effektstärke d. Nach Cohen (1988) gilt ein Effekt als groß, wenn $d \geq 0.80$ ist. Die Anzahl dieser Fälle ist in Tabelle 4 ebenfalls dargestellt. Von den jeweils 19 t-Tests fielen demnach 2 – 6 signifikant aus und von den signifikanten Mittelwertsunterschieden sind bis zu 3 pro Klasse auch groß im Sinne von Cohen. Nimmt man alles in allem, so muss man doch zumindest gelegentlich mit echten Unterschieden in der Schwierigkeit der Diktate rechnen, obgleich die zu diktierenden Wörter jeweils per Zufall aus der Grundmenge gezogen wurden und obgleich die Variationen um die Regressionslinien relativ gering sind. Für eine Verlaufsdiagnostik stellt das Ergebnis dennoch eine Beeinträchtigung dar.

Häufigkeitsverteilung der individuellen Verlaufskurven

Die Sichtung der vorliegenden 121 Datensätze aus den sechs Klassen führte zur Unterscheidung von vier Mustern von Verlaufskurven. *Linear ansteigende Verläufe* belegen eine der Tendenz nach stetige, oft jedoch durch starke Schwankungen gekennzeichnete Leistungsverbesserung. Solche Verläufe wird man am ehesten erwarten. *Verläufe mit Ceilingeffekten* sind unvermeidlich, wenn man an die obere

Grenze der geforderten Leistungen stößt. Kinder mit diesen Verläufen beherrschen den Grundwortschatz entweder von Anfang an oder im Laufe der Übungen.

Verläufe, die keine Verbesserung erkennen lassen, wird man allerdings weniger erwarten. Solche Verläufe bleiben unter oft starken Schwankungen auf ihrem Ausgangsniveau stehen. In einigen Fällen kommen sogar (meist lineare) Verschlechterungen vor. *Nichtlineare Verläufe* zeigen mitunter ein periodisches Auf und Ab. Den Verläufen kann eine Leistungsverbesserung zugrunde liegen, es kommt aber auch vor, dass sie eine solche nicht erkennen lassen.

Wie häufig sind die genannten Verläufe im vorliegenden Material anzutreffen? Eine Auskunft hierzu bietet Tabelle 5. Strathmann und Klauer (2008) hatten in einer früheren Studie unter anderem drei Sonderschulklassen der Mittel- und Oberstufe einbezogen, die leistungsmäßig dem Niveau der hier vorgestellten Grundschulklassen in etwa entsprachen. Die Verteilung der Lernkurven dieser Kinder auf die vier Varianten ist in Tabelle 5 zum Vergleich mit einbezogen. Da es hier keine statistisch bedeutsamen Unterschiede weder zwischen den sechs Klassen der beiden Grundschulen noch zwischen den drei Klassen der Sonderschule gibt, sei in der Tabelle nur zwischen den beiden Schulformen differenziert.

Wie man sieht, sind linear ansteigende Verläufe in den Stichproben beider Schulformen gleich häufig vertreten. Allerdings gibt es in der Sonderschule kaum Verläufe mit Deckeneffekt: Nur ein Kind hat eine Leistungsentwicklung genommen, die für eine Beherrschung des Grundwortschatzes spricht. Aber selbst in den Grundschulklassen ist der Anteil dieser Kinder relativ gering. Auffällig erscheint in beiden Schulformen der Anteil der Kinder, die in der Beobachtungszeit offenbar keinen Lernzuwachs in der Beherrschung des Grundwortschatzes erkennen lassen.

Diskussion

Methodologische Aspekte

Die *Reliabilität* des Verfahrens wurde schon wiederholt von anderen Autoren mit unterschiedlichen Inhalten und verschiedenen Gruppen von Lernenden überprüft. Bei der Lernfortschrittsmessung werden üblicherweise die nacheinander erhobenen unterschiedlichen Tests miteinander korreliert, so dass man von der Paralleltestreliabilität spricht. Dabei wurden durchweg zufrieden stellende Werte sowohl von amerikanischen Forschern berichtet (vgl. die Übersicht in Klauer, 2006, und in Walter, 2008), aber auch von deutschen Autoren. Unlängst haben nämlich

Tab. 5: Häufigkeitsverteilung der Verlaufstypen in Prozent

Schulform	Linear ansteigend	Mit Ceiling-effekt	Keine Verbesserung	Nichtlineare Varianten
Grundschule (N = 121)	42 %	16 %	30 %	12 %
Sonderschule (N = 37)	43 %	4 %	43 %	10 %

Walter (2008; 2009a) sowie Strathmann und Klauer (2008) umfangreiche Daten zur Paralleltestreliabilität und zur Validität der Lernfortschrittsmessung vorgelegt, die im Fall von Walter und für das Lesen als sehr befriedigend angesehen werden können. Die in der vorliegenden Studie gefundenen mittleren Reliabilitäten entsprechen allerdings keineswegs den Erwartungen, wie sie in Hypothese 1 und in Anlehnung an die vorliegende Literatur formuliert worden waren. Die Ergebnisse sind zweifellos unbefriedigend, und zwar unabhängig von der Methode der Reliabilitätsschätzung. Bemerkenswert sind allerdings die Paralleltestreliabilitäten unmittelbar aufeinander folgender Diktate. Möglicherweise variieren die Leistungen der Kinder im Laufe der Zeit doch stärker als vermutet.

Günstiger stellen sich die Befunde hinsichtlich der Validität des Verfahrens gemäß der Hypothesen 2 und 3 dar. Deren Erwartungen konnten bestätigt werden.

Ein weiterer wichtiger Aspekt des Verfahrens bezieht sich auf die Homogenität der Schwierigkeit der Aufgabenstichproben, worauf sich Hypothese 4 bezog. Die durchgeführten Signifikanztests zeigten, dass man keinesfalls durchweg gleiche Schwierigkeiten der Aufgabenstichproben unterstellen kann, auch wenn meist nur relativ kleine Schwierigkeitsunterschiede festzustellen waren. Tatsächlich zeigen die Verlaufskurven der Klassen nur eine vergleichsweise geringe Variabilität um die Regressionslinien.

Dennoch sollte die Frage der Homogenität der Testschwierigkeiten in der Lernverlaufsdiagnostik ernst genommen werden, was unserer Kenntnis nach in der Literatur nicht hinreichend geschieht.

Zusammenfassend ist also festzuhalten, dass weder die Reliabilitätskoeffizienten noch die Tests auf Homogenität der Schwierigkeit der einzelnen Diktate zu wirklich befriedigenden Ergebnissen führten. Für die Zukunft wird zu prüfen sein, in welcher Weise die Unterschiede in den Schwierigkeiten der Testdiktate überwunden und die Testreliabilitäten verbessert werden können.

Eine Möglichkeit besteht darin, die Zahl der auszuwählenden Items zu erhöhen. Es ist immerhin denkbar, dass größere Aufgabenstichproben repräsentativer hinsichtlich der mittleren Schwierigkeit ausfallen. Vielleicht genügen 20 Wörter nicht aus einer Grundmenge von 1334 Wörtern, um eine auch in der Schwierigkeit repräsentative Stichprobe darzustellen. Legt man die Samplingtheorie und die Formel von Berk (1980, vgl. Klauer, 1987, S. 26) zugrunde, so sind 20 Items sicher zu wenig, wenn ein hohes Lehrziel angestrebt wird. Diese Überlegung widerspricht zwar einem wichtigen Anliegen der amerikanischen CBM, die auf kurze Tests zugunsten häufiger Erhebungen setzt, was uns veranlasst hatte, nicht mehr als 20 Wörter zu diktieren. Möglicherweise kommt man aber nicht umhin, im Fall der Rechtschreibung deutlich größere Stichproben zu wählen.

Eine zweite Möglichkeit besteht darin, die Grundmenge von Wörtern in Teilmengen zu zerlegen, um dann stratifiziert – zufällige Stichproben zu ziehen, also die Teilmengen nach einem vorher festgelegten Schlüssel in jeder Zufallsstichprobe zu berücksichtigen (vgl. Klauer, 1987, S. 24 f). Wir haben hier die Teilmengen der Wortarten, die ihrerseits unterteilt sind in nicht flektierte und flektierte Formen. Ferner unterscheiden sich die Wörter nach der Anzahl ihrer Silben, und es ist bekannt, dass Wörter mit weniger Silben rascher beherrscht werden, im Lesen wie im Schreiben. Denkbar wäre also, die Grundmenge klar zu segmentieren und dann

die Stichproben so zu ziehen, dass eine jede die Teilmengen in dem Verhältnis bringt, wie es in der Grundmenge gegeben ist.

Schließlich war oben schon darauf verwiesen worden, dass Reimwörter partiell identisch sind („lachten" und „brachten"). Der Anteil solcher sich reimender Wörter dürfte die Schwierigkeit einer Aufgabenstichprobe ebenfalls beeinflussen. Möglicherweise müsste auch dieser Aspekt kontrolliert werden.

Soweit zur Verbesserung der Itemstichproben. Das hier eingesetzte Verfahren bietet darüber hinaus interessante Möglichkeiten sowohl zur Messung des Lernfortschritts im Laufe der Zeit als auch zur Prognose zukünftigen Lerngewinns. Bei linearen Regressionen, wie sie hier in den Klassenwerten gegeben sind, stellt der Steigungskoeffizient ein *Maß für den Lernfortschritt* im Verlauf der Zeit dar. Der Lernzuwachs ist es aber, der für Lehrpersonen besonders wichtig sein sollte, und dafür fehlte bislang ein geeignetes Maß. In den USA gibt es konsequenterweise verstärkt die Tendenz, die Bewertung von Schulen und Lehrpersonen nicht vom erreichten Leistungsstand abhängig zu machen, sondern vom Lern*zuwachs*, den sie vermitteln konnten. Für dieses „value – added assessment" (Raudenbush, 2004) eignet sich die Lernverlaufsdiagnostik besonders.

Lineare Regressionen ermöglichen darüber hinaus *Voraussagen in die Zukunft*. Unter der Annahme, dass sich die Lernbereitschaft und die (wesentlich von der Lehrkraft gestalteten) Lernbedingungen nicht erheblich verändern, lässt sich die weitere Entwicklung für die überschaubare Zukunft voraussagen. Im Fall etwa von Klasse 2 in Figur 2 beträgt der Regressionskoeffizient laut Tabelle 3 genau b = 0,25, was bedeutet, dass innerhalb von etwa 4 Wochen durchschnittlich ein weiteres Wort schriftsprachlich beherrscht wurde. Gemäß diesem Trend lässt sich vorhersagen, dass nach etwa 40 Schulwochen, also nach ungefähr einem Schuljahr die Regressionsgerade um 10 Wörter nach oben verschoben sein müsste und dann etwa bei 19 richtigen von den 20 diktierten Wörtern enden würde, weil die Klasse am Ende der 2. Klasse einen Mittelwert von knapp 9 erreicht hatte. Eine solche Voraussage unterstellt lediglich gleichen Lernfortschritt in der Zukunft. Solche Voraussagen sind zwar gut begründet, aber keineswegs üblich. Voraussagen lassen sich später überprüfen, so dass sich hier eine Möglichkeit eröffnet, den Lernfortschritt von Gruppen oder einzelnen Kindern nachträglich einzuschätzen und zu bewerten. Es wäre schon von Bedeutung beurteilen zu können, ob eine Klasse in diesem Zeitabschnitt so viel an Kompetenzzuwachs erfahren hat, wie dies im vorhergegangenen Zeitabschnitt (etwa bei der früheren Lehrkraft) der Fall war.

Pädagogische Aspekte

Was nun die Entwicklung der Rechtschreibkompetenz betrifft, so erlernen deutschsprachige Kinder die Rechtschreibung relativ langsam, auch wenn es sich nicht um lese-rechtschreibschwache Kinder handelt (Klicpera & Gasteiger-Klicpera, 1995; Schneider & Stefanek, 2007). Es mag einsichtige Gründe geben, wenn sich einzelne Kinder vorübergehend in der Rechtschreibung sogar verschlechtern. Solche Verläufe sind in dem vorliegenden Datensatz allerdings nur äußerst selten vorgekommen, so dass es sich wahrscheinlich um Bedingungen handelt, die in den Einzelfällen wirksam waren. Dass

aber fast ein Drittel der Grundschüler in dem Beobachtungszeitraum praktisch keine Fortschritte in der Rechtschreibung gemacht hat – in der Sonderschule reicht der Wert fast bis zur Hälfte der Kinder –, dieses Ergebnis hätte man wohl kaum vermutet. Der vielfach dokumentierte langsame Lernfortschritt in der Rechtschreibung hängt möglicherweise auch mit diesem Faktum zusammen: Selbst wenn weitaus die meisten der Schülerinnen und Schüler gute Lernfortschritte machen, so werden die Mittelwerte der ganzen Klassen nur langsam ansteigen, wenn ein beachtlicher Anteil von Kindern keine oder nur sehr geringe Zuwächse zu verzeichnen hat. Der Lernfortschritt der meisten Kinder wird so durch den Klassenmittelwert eindeutig unterschätzt.

Der fehlende Lernzuwachs vieler Kinder ist wohl nur dadurch erklärbar, dass die Lehrkräfte solche Entwicklungen nicht wahrnehmen. Es ist aber nicht hinnehmbar, wenn Lehrkräfte nicht bemerken, wie ein beachtlicher Anteil von Schülern keinerlei Fortschritt über längere Zeit erbringt. Hier könnte die Lernverlaufsdiagnostik entscheidende Hilfestellung leisten. Die wiederholte Leistungsmessung ist als solche zweifellos selbst schon ein Faktor, der den Lernzuwachs beeinflusst. Wenn es aber gelingt, den Lehrkräften Instrumente an die Hand zu geben, die es gestatten, auf einer einheitlichen Skala den Lernzuwachs während des Schuljahres regelmäßig und mit wenig Aufwand zu messen, so hätten die Lehrkräfte eine elegante Methode, den Lernfortschritt jedes einzelnen Kindes wie den der Klasse insgesamt zu dokumentieren. Die Möglichkeiten dafür zu schaffen wäre eine lohnende Aufgabe für Forschung und Entwicklung.

Literatur

Augst, G. (1989). Schriftwortschatz: Untersuchungen und Wortlisten zum orthographischen Lexikon bei Schülern und Erwachsenen. Frankfurt/M: Peter Lang.

Berk, R. A. (1980). Estimation of test length for domain-referenced reading comprehension tests. Journal of Experimental Education, 48, 188-193.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hilsdale: Erlbaum.

De Gruijter, D. N. M. & van der Kamp, L. J. T. (1984). Statistical methods in psychological and educational testing. Lisse: Swets & Zeitlinger.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. Exceptional Children, 52, 219-232.

Deno, S. L. (2003). Developments in curriculum-based measurement. The Journal of Special Education, 37, 184-192.

Diehl, K. & Hartke, B. (2007). Curriculumnahe Lernfortschrittsmessungen. Sonderpädagogik, 37, 195-211.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. School Psychology Review, 33, 188-192.

Fuchs, L. S. & Fuchs, D. (1993). Formative evaluation of academic progress: How much growth can we expect? School Psychology Review, 22, 1-30.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70.

Klauer, K. J. (1984). Über Parallelität, Reliabilität und Validität kontentvalider Paralleltests. Diagnostica, 30, 67-80.

Klauer, K. J. (1987). Kriteriumsorientierte Tests. Göttingen: Hogrefe.

Klauer, K. J. (1968). Orthographisches Lernen als Funktion der Lehrmethode, des Leistungs- und des Intelligenzniveaus. Zeitschrift für erziehungswissenschaftliche Forschung, 2, 39-54.

Klauer, K. J. (2005). Das Experiment in der pädagogischen Forschung. Münster: Waxmann (2. Aufl.).

Klauer, K. J. (2006). Erfassung des Lernfortschritts durch curriculumbasierte Messung. Heilpädagogische Forschung, 32, 16-26.

Klauer, K. J. & Dänecke, K. (1981). Wie parallel sind lehrzielvalide Paralleltests? Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 13, 181-189.

Klicpera, C. & Gasteiger-Klicpera, B. (1995). Psychologie der Lese- und Schreibschwierigkeiten. Entwicklung, Ursachen, Förderung. Weinheim: Beltz.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Naumann, C. L. (1999). Orientierungswortschatz. Die wichtigsten Wörter und Regeln für die Rechtschreibung Klasse 1 bis 6. Weinheim: Beltz (4. Aufl.).

Plickat, H. H. (1987). Deutscher Grundwortschatz. Weinheim: Beltz (3. Auflage).

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? Journal of Educational and Behavioral Statistics, 29, 121-129.

Schneider, W. & Stefanek, J. (2007). Entwicklung der Rechtschreibleistung vom frühen Schul- bis zum frühen Erwachsenenalter. Zeitschrift für Pädagogische Psychologie, 28, 77-82.

Schrader, F.-W. (2001). Diagnostische Kompetenz von Eltern und Lehrern. In D. H. Rost (Hrsg.), Handwörterbuch Pädagogische Psychologie (S.95-100). Weinheim: BeltzPVU.

Strathmann, A. & Klauer, K. J. (2008). Diagnostik des Lernverlaufs. Eine Pilotstudie am Beispiel der Entwicklung der Rechtschreibkompetenz. Sonderpädagogik, 38, 5-24.

VanDerHeyden, A. M., Witt, J. C., Naquin, G. & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. School Psychology Review, 30, 363-382.

Walter, J. (2008). Curriculumbasiertes Messen (CBM) als lernprozessbegleitende Diagnostik: Erste deutschsprachige Ergebnisse zur Validität, Reliabilität und Veränderungssensibilität eines robusten Indikators zur Lernfortschrittsmessung. Heilpädagogische Forschung, 34, 62-79.

Walter, J. (2009a). Eignet sich die Messtechnik „MAZE“ zur Erfassung von Lesekompetenzen als lernprozessbegleitende Diagnostik? Heilpädagogische Forschung, 35, 62-75.

Walter, J. (2009b). Theorie und Praxis Curriculumbasierten Messens (CBM) in Unterricht und Förderung. Zeitschrift für Heilpädagogik, 60, 162-170.

Anhang: Beispiel einer Wortliste

1 Schlösser	11 bezahlst
2 gelacht	12 bekommt
3 sonst	13 Hunde
4 gebacken	14 gebadet
5 kaufen	15 eigentlich
6 heute	16 morgens
7 verletzte	17 leiser
8 enger	18 darauf
9 baust	19 am wichtigsten
10 bewegt	20 am kleinsten

Anschriften der Autoren:

Prof. Dr. Alfons Strathmann
Humanwissenschaftliche Fakultät
Universität zu Köln
Klosterstr. 79
50931 Köln
alfons.strathmann@uni-koeln.de

Prof. Dr. Karl Josef Klauer
Robert-Stolz-Weg 15
42781 Haan
klauerk@uni-duesseldorf.de

Dr. Michaela Greisbach
Humanwissenschaftliche Fakultät
Universität zu Köln
Klosterstr. 79
50931 Köln