Marcus Hasselhorn · Wolfgang Schneider
Ulrich Trautwein (Hrsg.)

Lernverlaufs-diagnostik

Tests und Trends
N. F. Band 12
Jahrbuch der pädagogisch-psychologischen Diagnostik

HOGREFE

Lernverlaufsdiagnostik

Jahrbuch der pädagogisch-psychologischen Diagnostik

Tests und Trends

Neue Folge Band 12

Lernverlaufsdiagnostik

hrsg. von Prof. Dr. Marcus Hasselhorn, Prof. Dr. Wolfgang Schneider und Prof. Dr. Ulrich Trautwein

---

Herausgeber der Reihe:
Prof. Dr. Marcus Hasselhorn, Prof. Dr. Wolfgang Schneider,
Prof. Dr. Ulrich Trautwein

Lernverlaufs-diagnostik

herausgegeben von

Marcus Hasselhorn, Wolfgang Schneider
und Ulrich Trautwein

HOGREFE

GÖTTINGEN · BERN · WIEN · PARIS · OXFORD · PRAG
TORONTO · BOSTON · AMSTERDAM · KOPENHAGEN
STOCKHOLM · FLORENZ · HELSINKI

Prof. Dr. Marcus Hasselhorn, geb. 1957. 1977–1983 Studium der Psychologie und Pädagogik. 1986 Promotion. 1993 Habilitation. 1993–1997 Professor für Entwicklungspsychologie an der TU Dresden. 1997–2007 Leiter der Abteilung Pädagogische Psychologie und Entwicklungspsychologie an der Universität Göttingen. Seit 2007 Leiter der Arbeitseinheit Bildung und Entwicklung am Deutschen Institut für Internationale Pädagogische Forschung (DIPF) in Frankfurt am Main.

Prof. Dr. Wolfgang Schneider, geb. 1950. 1969–1975 Studium der Psychologie, Theologie und Philosophie. 1976–1981 Wissenschaftlicher Mitarbeiter am Psychologischen Institut der Universität Heidelberg. 1979 Promotion. 1981–1982 Visiting Scholar an der Stanford University (USA). 1982–1991 Wissenschaftlicher Mitarbeiter am Max-Planck-Institut für psychologische Forschung in München. 1988 Habilitation. 1990–1991 Vertretung und seit 1991 Inhaber des Lehrstuhls für Pädagogische und Entwicklungspsychologie an der Universität Würzburg.

Prof. Dr. Ulrich Trautwein, geb. 1972. 1992–1999 Studium der Psychologie. 1999 Diplom in Psychologie. 2002 Promotion. 2005 Habilitation. Seit 2008 Universitätsprofessor für Empirische Bildungsforschung an der Universität Tübingen.

Göttingen · Bern · Wien · Paris · Oxford · Prag · Toronto · Boston
Amsterdam · Kopenhagen · Stockholm · Florenz · Helsinki
Merkelstraße 3, 37085 Göttingen

http://www.hogrefe.de
Aktuelle Informationen · Weitere Titel zum Thema · Ergänzende Materialien

---

Satz: ARThür Grafik-Design & Kunst, Weimar
Format: PDF

ISBN 978-3-8409-2614-3

Nutzungsbedingungen:

Der Erwerber erhält ein einfaches und nicht übertragbares Nutzungsrecht, das ihn zum privaten Gebrauch des E-Books und all der dazugehörigen Dateien berechtigt.

Der Inhalt dieses E-Books darf von dem Kunden vorbehaltlich abweichender zwingender gesetzlicher Regeln weder inhaltlich noch redaktionell verändert werden. Insbesondere darf er Urheberrechtsvermerke, Markenzeichen, digitale Wasserzeichen und andere Rechtsvorbehalte im abgerufenen Inhalt nicht entfernen.

Der Nutzer ist nicht berechtigt, das E-Book – auch nicht auszugsweise – anderen Personen zugänglich zu machen, insbesondere es weiterzuleiten, zu verleihen oder zu vermieten.

Das entgeltliche oder unentgeltliche Einstellen des E-Books ins Internet oder in andere Netzwerke, der Weiterverkauf und/oder jede Art der Nutzung zu kommerziellen Zwecken sind nicht zulässig.

Das Anfertigen von Vervielfältigungen, das Ausdrucken oder Speichern auf anderen Wiedergabegeräten ist nur für den persönlichen Gebrauch gestattet. Dritten darf dadurch kein Zugang ermöglicht werden.

Die Übernahme des gesamten E-Books in eine eigene Print- und/oder Online-Publikation ist nicht gestattet. Die Inhalte des E-Books dürfen nur zu privaten Zwecken und nur auszugsweise kopiert werden.

Diese Bestimmungen gelten gegebenenfalls auch für zum E-Book gehörende Audiodateien.

Anmerkung:

Sofern der Printausgabe eine CD-ROM beigefügt ist, sind die Materialien/Arbeitsblätter, die sich darauf befinden, bereits Bestandteil dieses E-Books.

Inhaltsverzeichnis

Vorwort der Herausgeber

Seit dem Beginn der Neuen Folge der Reihe „Tests und Trends – Jahrbuch der pädagogisch-psychologischen Diagnostik“ im Jahr 2000 wurde in den nun 14 Jahren ihres Bestehens immer wieder versucht, den neuesten Stand diagnostischer Möglichkeiten in unterschiedlichen Inhaltsbereichen schulischen Lernens zu dokumentieren. In den bislang erschienenen Bänden der Reihe wurden praktisch relevante Forschungsansätze und -ergebnisse zu spezifischen Lernleistungen, spezifischen Lernvoraussetzungen sowie zu lernbegleitenden Fähigkeiten, Funktionen und Dispositionen unter diagnostischem Blickwinkel berichtet. Im Mittelpunkt der Betrachtung standen dabei meist standardisierte Testverfahren zur Erfassung von Kompetenzen in ausgewählten Inhaltsbereichen, die als „Statustests“ das Leistungsvermögen von Schülerinnen und Schülern zu einem bestimmten Zeitpunkt erfassen. Diese Art von Bestandsaufnahme wird auch als „summative Diagnostik“ charakterisiert und bezeichnet in der Regel die Evaluation des Ergebnisses eines langfristigen Lernvorgangs.

Der hier vorliegende 12. Band der Neuen Folge beschäftigt sich im Unterschied dazu mit Möglichkeiten der „formativen Leistungsdiagnostik“, die schon seit mehreren Jahrzehnten in den USA relativ populär ist und in den letzten Jahren auch verstärkt im deutschsprachigen Raum Beachtung gefunden hat. Wenn auch die Definition dieses Konzepts in der Literatur nicht ganz eindeutig ausfällt, so ist hier meist eine systematische *Lernverlaufsdiagnostik* gemeint, die Informationen über die Entwicklung von Schülerleistungen gibt und von der Grundannahme geleitet wird, dass die beständige Rückmeldung von Leistungsveränderungen an Schülerinnen und Schüler sowie Lehrerinnen und Lehrer zu insgesamt besseren Ergebnissen führt.

Der vorliegende Band gibt den Stand der neueren deutschsprachigen Entwicklungsarbeit im Bereich der *formativen* Diagnostik und Evaluation wieder. Das einleitende Kapitel von *Karl Josef Klauer* erläutert die begriffliche wie auch methodische Problematik und führt in die Konzepte der curriculumsbasierten Messung und des Ansatzes der „Response to Intervention“ (RTI) ein, die in den Kapiteln des Bandes eine wichtige Rolle spielen. Das sich anschließende Kapitel von *Uwe Maier* gibt einen Überblick über theoretische Grundannahmen und Methoden der formativen Leistungsdiagnostik in der Sekundarstufe und illustriert Anwendungsmöglichkeiten im mathematisch-naturwissenschaftlichen wie auch im sprachlichen Unterricht, die auch Ansätze computergestützter formativer Leistungsdiagnostik mit einschließen. In ähnlicher Weise überblicksorientiert ist das Kapitel zur formativen Leistungsbeurteilung von Anika *Bürgermeister, Eckhard Klieme, Katrin Rakoczy, Birgit Harks und Werner Blum*, das ebenfalls den Unterricht der Sekundarstufe in den Blick nimmt und die Möglichkeiten dieses Ansatzes im Rahmen eines DFG-geförderten Projekts zur Leistungsbeurteilung im deutschen Mathematikunterricht illustriert.

Der Beitrag von *Yvonne Blumenthal, Kristin Kuhlmann und Bodo Hartke* kontrastiert den klassischen „aptitude-treatment-interaction"-Ansatz (ATI) mit dem oben erwähnten RTI-Ansatz, um die Möglichkeiten unterschiedlicher Lösungsansätze für das Problem der optimalen Anpassung von Unterrichtsmethoden an unterschiedliche Lernvoraussetzungen von Schülern darzustellen. Das übergeordnete Ziel des RTI-Ansatzes, dessen Vorteil in einem gestuften Fördersystem liegt, wird in der Früherkennung und Prävention von Lernschwierigkeiten und Lernstörungen gesehen, was als sinnvolle Alternative zum ATI-Ansatz gelten kann. Das sich anschließende Kapitel von *Stefan Voß und Bodo Hartke* bezieht sich ebenfalls auf den RTI-Ansatz und verdeutlicht im Rahmen eines Literaturüberblicks, dass hier curriculumsbasierte Messverfahren (CBM) für die Erfassung des Lernverlaufs in unterschiedlichen Inhaltsbereichen von zentraler Bedeutung sind. Ein aktuelles Anwendungsbeispiel für den Einsatz von RTI bietet der Beitrag von *Kathrin Mahlau, Yvonne Blumenthal, Kirsten Diehl, Anne Schöning, Simon Sikora, Stefan Voß* und *Bodo Hartke*, der das „Rügener Inklusionsmodell" beschreibt und aufzeigt, wie Mehrebenenprävention funktionieren kann, wenn je nach Ausmaß des Lerndefizits in den Bereichen Deutsch und Mathematik, aber auch im Bereich emotionaler und sozialer Entwicklung unterschiedliche Förderebenen (regulärer Unterricht, Kleingruppenförderung, individuelle Förderung) aktiviert werden.

Kerstin Bäuerlein, Frank Niklas und Wolfgang Schneider stellen mit dem Verfahren „Fähigkeitsindikatoren Primarschule" (FIPS) ein computergestütztes Diagnoseinstrument für den Einsatz in der ersten Schulklasse vor. Über den Vergleich der Lernausgangslage in den Bereichen Wortschatz, Lautbewusstheit, Lesen und Mathematik zu Beginn der Schulzeit mit den im Verlauf des ersten Schuljahrs erzielten Lernfortschritten in diesen Bereichen können wichtige Informationen für Lehrkräfte bereitgestellt werden, die adaptive Fördermaßnahmen zu einem frühen Zeitpunkt ermöglichen. *Kirsten Diehl* stellt mit dem Inventar zur Erfassung der Lesekompetenzen von Erstklässlern (IEL-1) ein formatives Evaluationsverfahren zur Erfassung der Lernfortschritte im Lesen vor, das ähnlich wie FIPS die Schuleingangsphase im Blick hat, dabei aber auf den Leselernvorgang fokussiert und in der formativen Diagnostik so kleinschrittig angelegt ist, dass Rückstände im Leselernprozess frühzeitig erkannt werden können. Möglichkeiten der Lernfortschrittsdiagnostik und der Verlaufsdiagnostik im Bereich Lesen werden auch umfassend im Kapitel von *Jürgen Walter* am Beispiel von zwei diagnostischen Verfahren beschrieben, die sorgfältig konzipiert und evaluiert wurden und im Grundschulbereich wie auch in der Sekundarstufe einsetzbar sind. Ermutigende Befunde zur Lernverlaufsdiagnostik Mathematik für die zweite bis vierte Klassenstufe der Grundschule liefert der Beitrag von *Alfons M. Strathmann*, in dem ein Verfahren vorgestellt wird, dass mittels CD die Generierung von immer neuen Einzeltests möglich macht. Lernverlaufskurven lassen sich damit problemlos für jedes Kind einer Klasse wie auch die Klasse als Ganzes erstellen.

Zwei Beiträge der Münsteraner Arbeitsgruppe um Elmar Souvignier setzen sich mit der Wirksamkeit formativer Diagnostik und der Möglichkeit von internetba-

sierten Einsätzen auseinander. *Elmar Souvignier, Natalie Förster* und *Elisabeth Schulte* gehen davon aus, dass idealerweise diagnostische Verfahren mit hoher psychometrischer Güte wiederholt zum Einsatz kommen müssen, um die Wirksamkeit des Verfahrens zu optimieren. Sie stellen mögliche Untersuchungsdesigns bei der Wirksamkeitsprüfung formativer Assessments vor und präsentieren Befunde ihres Forschungsprogramms zur Evaluation formativer Diagnostik. Das sich anschließende Kapitel von *Elmar Souvignier, Natalie Förster* und *Martin Salaschek* geht weiter ins Detail und illustriert die Möglichkeiten internetbasierter Lernverlaufsdiagnostik in den Bereichen Lesen und Mathematik am Beispiel von „quop", einem internetbasierten System, das eine ökonomische Durchführung der formativen Diagnostik im Regelunterricht der Grundschule sowie eine automatisierte Auswertung und Dokumentation der Ergebnisse ermöglicht.

Der Beitrag von *Peter May, Jasmine Bennöhr und Carina Berger* behandelt die Hamburger Testserie „KEKS", ein Verfahren zur „Kompetenzerfassung in Kindergarten und Schule", das Kernkompetenzen in den Bereichen Deutsch, Mathematik, Englisch und – bei Kindern mit Migrationshintergrund – unterschiedlichen Herkunftssprachen erfasst. KEKS erfüllt die üblichen Gütekriterien und ist sowohl bei Kindern als auch bei Jugendlichen einsetzbar, umfasst dabei den Altersbereich ab 4 bis ca. 16 Jahre und kann demnach die individuelle Lernentwicklung über einen größeren Zeitraum abbilden. Im abschließenden Kapitel von *Jürgen Wilbert* geht es um methodische Herausforderungen und Besonderheiten der Lernverlaufsdiagnostik und die Illustration messtheoretischer Voraussetzungen, die bei Instrumenten zur Lernverlaufsdiagnostik erfüllt sein sollten, wenn sie sich in Forschung und Praxis der empirischen Bildungsforschung, der Sonderpädagogik und der Pädagogischen Psychologie fest etablieren wollen.

Insgesamt gesehen bietet der vorliegende Band einen umfassenden Überblick über neuere Entwicklungen im Bereich der Lernverlaufsdiagnostik, die das Potenzial dieses diagnostischen Ansatzes für die Erfassung der Leistungsentwicklung in unterschiedlichen schulischen Inhaltsbereichen und für unterschiedliche Altersgruppen illustrieren, gleichzeitig aber auch die Voraussetzungen für einen effektiven Einsatz deutlich machen.

Frankfurt, Würzburg, Tübingen, im Dezember 2013

Marcus Hasselhorn
Wolfgang Schneider
und Ulrich Trautwein

Kapitel 1

Formative Leistungsdiagnostik: Historischer Hintergrund und Weiterentwicklung zur Lernverlaufsdiagnostik

Karl Josef Klauer

Zusammenfassung

Die Unterscheidung von formativer und summativer Diagnostik wurde 1967 von Scriven eingeführt. Formative Diagnostik bezog sich auf wiederholte Messungen im Laufe eines Lernprozesses, summative Diagnostik auf die Messung des am Ende erreichten Leistungsstandes. Allerdings erhielt formative Diagnostik in den USA immer wieder neue Interpretationen, was zu vielfach beklagter Konfusion führte. Dennoch resultierten Ansätze, die sich als zukunftsträchtig erweisen sollten, so etwa das „Curriculum Based Measurement“ oder die „Response to Intervention“. Eine deutsche Entwicklung formativer Evaluation stellt die Lernverlaufsdiagnostik dar. Die Lernverlaufsdiagnostik erfordert (1) eine klare Definition der wiederholt zu erfassenden Kompetenz, (2) homogen schwierige, (3) änderungssensible Tests und (4) unter Umständen eine Abkehr von der klassischen Testtheorie.

1.1 Historische Entwicklung in den USA

1.1.1 Unterscheidung formativer und summativer Diagnostik

Die Unterscheidung von formativer und summativer Diagnostik geht auf Scriven (1967) zurück. Er bezeichnete formative Evaluation als die in Abständen wiederholte Evaluation eines sich im Vollzug befindlichen Lernprozesses und im Gegensatz dazu summative Evaluation als Evaluation des Ergebnisses am Schluss einer solchen Lernperiode oder eines Lernprogramms. Bloom (1969) griff die Unterscheidung auf und modifizierte sie deutlich. Formative Diagnostik sollte danach nicht nur den Lernfortschritt im Laufe des Prozesses *dokumentieren*, sondern zugleich auch den Prozess *fördern*, das Lernen *verbessern*, und zwar durch *Rückmeldung* der Lernergebnisse an die Lernenden einerseits und an die Lehrenden andererseits. In ihrem auch in Deutschland weit verbreiteten Handbuch sorgten Bloom, Hastings und Madaus (1971) für die allgemeine Verbreitung dieser Unterscheidung von formativer und summativer Evaluation und für die von Bloom vorgeschlagene Interpretation.

Dabei ist die Entwicklung aber nicht stehen geblieben. In der Folge wurde die Unterscheidung Allgemeingut, wobei allerdings viele Autoren die Begriffe mehr oder minder nach eigenem Gutdünken interpretierten. Tatsächlich wurde sehr bald schon eine große Konfusion um die Unterscheidung von „formativ“ und „summativ“ festgestellt, weil sie zwar allgemein akzeptiert und gebraucht, aber vielfach beliebig umgedeutet wurde (Misanchuk, 1978). Das betraf insbesondere die formative Evaluation, die ja auch das eigentlich Neue darstellte, denn die summative Evaluation bezieht sich im Wesentlichen auf die gängigen Schulleistungstests, wie sie etwa am Ende eines Schuljahres eingesetzt werden.

In besonders einflussreichen Abhandlungen hatten Black und Wiliam (1998; 2003) formative Evaluation im Anschluss an Bloom nicht als ein Verfahren definiert, das durch bestimmte *Eigenschaften* gekennzeichnet sei, sondern durch den *Zweck*, dem es dient, nämlich durch die Lernförderung. Ihr wichtigstes Argument bezogen sie aus einer Metaanalyse, die die bekannten Sonderpädagogen Fuchs und Fuchs (1986) veröffentlicht hatten und die glaubten nachzuweisen, dass formative Evaluation das Lernen in erheblichem Ausmaße voranbringt. Black und Wiliam hielten deshalb für erwiesen, dass formative Evaluation das Lernen fördert. Inzwischen gibt es aber erhebliche und gut begründete Zweifel an der Qualität dieser Metaanalyse und an den daraus gezogenen Schlussfolgerungen, wie Dunn und Mulvenon (2009) in einer elektronischen Zeitschrift differenziert belegen. Allerdings gilt es, hier zu unterscheiden: Tatsächlich gibt es gut begründete Nachweise, dass die regelmäßige Leistungsmessung mit entsprechenden Rückmeldungen die Leistung deutlich verbessern kann, nur ist dies eben nicht unter allen Bedingungen der Fall.

Die allgemein große Zustimmung, die die formative Diagnostik inzwischen erhalten hatte und noch erhält, führte immer mehr Autoren dazu, eigene Konzepte „formativer“, also lernfördernder Evaluation zu entwickeln. Das sollen einige Beispiele illustrieren. Beer und Bloomer (1986) traten für eine Kombination formativer *und* summativer Evaluation ein, wobei die formative Evaluation weitgehend qualitative Methoden verwenden sollte. Oder für Sadler (1989) ging es in der formativen Evaluation primär darum, das „Feedback“ auf differenziertere Weise zu vermitteln. Ferner definierten Nicol und MacFarlane-Dick (2006) formative Evaluation durch die Optimierung des *selbstregulierten* Lernens. Oder Garrison und Ehringhaus (2007) stellten fest, dass formatives Assessment durch aktive Einbeziehung der *Lernenden*, durch *Praxis* und noch durch manches andere gekennzeichnet sei. Am Beispiel der Fahrschule wollten sie klar machen, was sie meinten: Die Fahrprüfung sei summativ, aber die Fahrübungen vorher seien formativ. Wininger (2005) schlug sogar vor, formativ-summative Tests einzusetzen, also beide Varianten miteinander kombiniert zu verwenden, die von anderen Autoren als gegensätzlich eingeschätzt worden waren. Misanchuk (1978) war aber Jahre zuvor schon einen Schritt weiter gegangen und hatte eine dritte Variante eingeführt: Formative, summative und konfirmative Evaluation. Wesentlich später waren Dunn und Mulvenon (2009) in der Lage, noch weiter zu diffe-

renzieren und genau vier Varianten einzuführen. Sie unterschieden nun formatives Assessment von summativem Assessment und analog formative Evaluation von summativer Evaluation. Für die Autoren war die nahezu chaotische Konfusion, die sich im englischsprachigen Raum um formativ und summativ sowie um Evaluation und Assessment entwickelt hatte, der Anlass, um gedankliche Ordnung zu stiften. Ob sie damit bei ihren Kollegen Erfolg haben werden, darf man füglich bezweifeln, denn das terminologische Durcheinander ist bislang nicht wesentlich besser geworden.

Ein Motiv für die *Weiterentwicklung* dieser Testtradition wurde in den USA durch das „No Child Left Behind"-Gesetz von 2002 ausgelöst. Schon im Rahmen der älteren *Accountability*-Politik waren die Schulen gehalten, dafür zu sorgen, dass alle Kinder angemessen gefördert würden. Das neue Gesetz von 2002 machte die Schulen nun aber unmittelbar *verantwortlich* für etwaige Misserfolge von Kindern. Seither sind die Schulen gehalten, dass 95 % der Schülerinnen und Schüler die Vorgaben des Bundeslandes im Lesen und in Mathematik erreichen, ab dem Schuljahr 2013/2014 sollen sogar 100 % der Kinder die Norm erfüllen. Das Gesetz belegte die Schulen mit entsprechend harten Sanktionen, die von der Entlassung von Lehrkräften bis zur Schließung ganzer Schulen reichen können (Deno, 2003a). Diese Gesetzgebung führte zu einer Fülle von Aktivitäten, um Leistungsschwächen frühzeitig zu erkennen und um das Lernen entschieden zu fördern. Eine wirklich neue und folgenreiche Entwicklung stammt aus dem Bereich der Sonderpädagogik, das „curriculum-based measurement", welches allerdings schon weit vor dem Jahre 2002 einsetzte, dann aber durch das Gesetz starken Auftrieb erhielt.

1.1.2 Curriculum-Based Measurement (CBM)

Der Ansatz wurde von Stanley Deno und seinen Mitarbeitern an der Universität Minnesota seit 1972 entwickelt (vgl. auch Deno, 1985, 2003a, b). Nicht zufällig entstand das Konzept in der Sonderpädagogik, waren doch leistungsschwache Kinder auch schon von der älteren Accountability-Politik in besonderem Maße betroffen und Lehrer entsprechend motiviert, solche Kinder frühzeitig zu erkennen und entsprechend zu fördern. Denos Begriff lässt sich mit curriculumbasierter Messung leicht ins Deutsche übertragen, würde dann aber falsche Assoziationen auslösen. Tatsächlich sind die deutschen Schulleistungstests typischerweise an den Lehrplänen soweit wie möglich orientiert, und es ist in aller Regel möglich, mithilfe eines Schulleistungstests summativ festzustellen, ob ein Kind die Leistungen erbringt, die für das entsprechende Schuljahr vorgegeben sind. In dem Sinne könnte man deutsche Schulleistungstests durchaus als curriculumbasiert bezeichnen.

Genau solche Leistungsmessungen sind aber nicht gemeint mit „Curriculum-Based Measurement". Hier geht es gerade nicht um den Einsatz der üblichen

standardisierten Schulleistungstests, sondern idealerweise um von der Lehrkraft entwickelte Leistungstests, die genau das testen sollen, was im *aktuellen Unterricht* durchgenommen worden ist. Mit CBM soll also ermittelt werden, wie gut das hier und heute (etwa diese Woche) behandelte Teilziel des Unterrichts von den einzelnen Kindern bewältigt worden ist. Diese Art der Messung ist natürlich nur von den Lehrpersonen durchführbar, weil nur sie genau wissen, was zuletzt durchgenommen worden ist. Ein weiterer Aspekt, der nicht direkt aus der Namensgebung hervorgeht, bezieht sich auf die *Häufigkeit* der „curriculumbasierten" Leistungsmessung. Tatsächlich geht es nicht nur darum, die Tests direkt auf den Lehrstoff zu beziehen, der in der Klasse behandelt worden ist, sondern auch darum, solche Leistungserhebungen in relativ kurzen Abständen zu wiederholen, um die Entwicklung des Lernens über einen *längeren Zeitraum* abbilden zu können. Zu diesem Zweck sollen die Tests vergleichsweise kurz sein und aus relativ leichten Aufgaben bestehen, die sich auch rasch auswerten lassen, so dass die Tests insgesamt wenig Zeit in Anspruch nehmen. Auf diese Weise können sie eben oft durchgeführt werden und engen die für den eigentlichen Unterricht zur Verfügung stehende Zeit nicht besonders ein.

Den Gesichtspunkt häufiger Leistungsmessungen haben *curriculumbasierte* und *formative* Evaluation gemeinsam. Von daher sollte man erwarten, dass es hier enge Überschneidungen gibt. Aber das ist erstaunlicherweise nicht der Fall. Während „formative assessment" zu der Zeit im „main stream" der pädagogisch-psychologischen Diagnostik florierte, fand die „curriculum based" Diagnostik kaum allgemeinere Beachtung, sondern blieb weitgehend und bis heute ein Nischenprodukt und vornehmlich konzentriert auf den Bereich der Sonderpädagogik. Es gab zwar da und dort gelegentliche Überschneidungen (vgl. etwa Fuchs & Fuchs, 1986, 1993 und Fuchs, 2004), aber die etwas randständige Position der Sonderpädagogik blieb hier wie auch vielfach anderswo unverändert. Das wird schon an der Terminologie deutlich: So spielen die damals in dem Zusammenhang dominierenden Begriffe „evaluation" und „assessment" beim CBM kaum eine Rolle, sondern sind durch „measurement" ersetzt, und was mit „curriculum based" gemeint war, wäre vermutlich in den USA besser verstanden worden, wenn man von einer speziellen Variante von „formative evaluation" gesprochen hätte wie dies Fuchs und Fuchs (1993) in einer schulpsychologischen Zeitschrift auch versucht haben. Immerhin deutet die Verwendung des Begriffs „measurement" auf einen gewissen Anspruch in methodologischer Hinsicht.

Wie kann „curriculumbasierte" Messung konkret durchgeführt werden? Zur Konstruktion solcher Tests unterschied Fuchs (2004) zwei Wege. Der *erste Weg* besteht darin, einen Aufgabentyp herauszusuchen, der repräsentativ für die fragliche Kompetenz ist, um die es geht, und möglichst viele der relevanten Teiloperationen beansprucht, die dabei gefordert sind. Stichproben solcher Aufgaben sollten stets das Gleiche messen, möglichst gleich schwer sein und hoch miteinander korrelieren, aber auch mit relevanten anderen Kompetenzen. Ein klassisches

Beispiel hierzu sind die Lesetests, die im Rahmen der CBM vielfach eingesetzt wurden. Dazu sucht man beispielsweise ein den Kindern nicht bekanntes Buch heraus, das ihrer Lesekompetenz angemessen ist, und lässt jedes Kind an einer beliebigen Stelle für genau eine Minute laut lesen, um zu registrieren, wie viele Wörter das Kind in der Zeit richtig liest. Diese Prozedur lässt sich, wenn man will, etwa jede Woche wiederholen, wobei immer nur an einer anderen Textstelle begonnen wird. Notiert man dann die Anzahl der jeweils in einer Minute richtig gelesenen Wörter, so erhält man mit der Zeit einen Entwicklungsverlauf, der die Fortschritte im Lesen schön dokumentieren lässt. So konnte wiederholt gezeigt werden, dass die Anzahl richtig gelesener Wörter von heute gut mit der Anzahl korreliert, die später gelesen wurde, aber auch gut mit dem Leseverständnis und mit manchen anderen Variablen.

Bei Lesetests wurde auch eine weitere Variante öfters eingesetzt, bei der es primär um das Leseverständnis statt um das formale Lesen geht, nämlich Lückentexte. Mitunter wurden Lückentexte sogar per PC geboten, beispielsweise ab dem zweiten Schuljahr (Fuchs & Fuchs, 1993). Der erste Satz wurde dabei vollständig geboten und ab dem zweiten Satz wurde systematisch jedes siebte Wort gestrichen. Dazu gab es dann drei Auswahloptionen, die möglichst gleich lang sein sollten und von denen zwei natürlich deutlich falsch waren. Die Kinder erhielten dann genau 2,5 Minuten Zeit, und registriert wurde die Anzahl richtiger Einsetzungen. Auch dieses Vorgehen bewährte sich durch den Nachweis entsprechend günstiger Korrelationen mit anderen Variablen.

Rechtschreibtests werden ganz analog eingesetzt. Man wählt eine Stichprobe von Wörtern, diktiert immer nur einzelne Wörter ohne jeden Kontext und stellt am Ende fest, wie viele Wörter dieses Mal richtig geschrieben worden sind.

Bemerkenswert an diesem ersten Weg ist allerdings der Umstand, dass von „curriculumbasierter“ Messung im Sinne der CBM eigentlich keine Rede mehr sein kann. Erfasst wird keineswegs das, was gerade im Unterricht behandelt wurde, wie dies hätte sein sollen. Tatsächlich konnten Riley-Heller, Kelly-Vance und Shriver (2005) nachweisen, dass es keinen Unterschied macht, ob man in der *Klasse* eingesetzte Texte oder Texte aus standardisierten *Schulleistungstests* verwendet. Es gibt auch andere Hinweise darauf, dass man es mit dem Verständnis von CBM im ursprünglichen Sinne dann doch nicht so genau nahm.

Der *zweite Weg* stellt sich etwas komplexer dar. Nach meinem Verständnis des Ansatzes versucht man zu präzisieren, was genau mit der jeweilig interessierenden Kompetenz gemeint ist – und zwar durch die Aufgabenmenge, die beherrscht werden soll. Das Verfahren kommt in der Mathematik, aber auch im Sachunterricht infrage. Ist die anzustrebende Kompetenz durch die Beherrschung einer Aufgabenmenge definiert, so kann man eine kleine Stichprobe von Aufgaben erstellen und als Test den Kindern vorlegen, um später eine andere Stichprobe

von Aufgaben zu wählen. Gezählt wird dann ebenfalls die Anzahl richtig gelöster Aufgaben.

Es gibt eine ganze Reihe von Studien, in denen die testtheoretischen Eigenschaften von Verfahren der „curriculumbasierten“ Leistungsmessung überprüft worden sind. Zweifellos sind die Verfahren objektiv und relativ zügig auswertbar, die Reliabilität erwies sich etwa im Sinne der Retestreliabilität als durchweg erfreulich hoch. Das gilt auch für die Validität, etwa die Korrelation mit standardisierten Tests (Good & Jefferson, 1998; Riley-Heller et al., 2005; VonDerHeyden, Witt, Naquin & Noell, 2001). Natürlich gibt es ebenso Fälle weniger erfreulicher Resultate – und die werden wohl auch nicht alle veröffentlicht worden sein.

1.1.3 Response to Intervention (RTI)

Eine jüngere Entwicklung im Bereich der amerikanischen Sonderpädagogik betrifft das Verfahren zur Identifizierung von Kindern mit „Learning Disability“. Definiert wurde diese Lernschwäche im Jahre 1977 als „underachievement“, d. h. durch die Diskrepanz zwischen Fähigkeit und tatsächlicher Leistung. Es geht also nicht um schlicht lernschwache Kinder, sondern um solche, die nicht das an Leistung erbringen, was sie intellektuell eigentlich erbringen könnten. Da die Vorgehensweise zur Identifizierung solcher Kinder unbefriedigende Ergebnisse brachte – inzwischen stieg die Anzahl so eingestufter Probanden um über 200 % (vgl. Bradley, Danielson & Doolittle, 2005) – wurden neue Möglichkeiten eingeführt. Dazu gehört das Verfahren der Reaktion auf gezielte Förderungen („Response to Intervention“), das bewusst im Kontrast zum Einsatz von Intelligenztests entwickelt wurde. Man unterscheidet dabei drei Niveaustufen (Bradley et al., 2005). Stufe eins bezieht sich auf das Verhalten und die Ergebnisse im regulären Unterricht. Stufe zwei und drei stellen sich als eine Art Quasi-Experimente dar: In Stufe zwei erhält eine Gruppe von Kindern, in Stufe drei erhalten einzelne Kinder speziell auf sie zugeschnittene und zeitlich begrenzte Förderung, wobei deren Effekt auf die geförderte Kompetenz systematisch erfasst und getestet wird. Wenn der Erfolg ausbleibt, findet erneut eine spezielle Förderung mit entsprechender Leistungsmessung statt, und wenn der Erfolg auch dann noch ausbleibt, wird erst über die Sonderschulbedürftigkeit des Kindes entschieden. So kommt es auch bei diesem Verfahren zu wiederholten Leistungsmessungen in kurzen Abständen, also zu einer weiteren Variante formativer Evaluation, deren Änderungssensibilität allerdings unabhängig und zuvor erwiesen sein sollte (vgl. hierzu Abschnitt 1.3.3).

RTI-Ansätze werden primär im Leistungsbereich eingesetzt, so etwa bei Leseproblemen (Fuchs & Fuchs, 2006), doch konnten Fairbanks, Sugai, Guardino und Lathrop (2007) in differenzierten Einzelfall-Studien zeigen, dass die Verfahren auch bei Problemen des Sozialverhaltens von Kindern effizient sein können.

1.2 Die Entwicklung im deutschsprachigen Raum

Die Unterscheidung von formativer und summativer Evaluation wurde von der deutschen Pädagogischen Psychologie frühzeitig aufgegriffen. Dies geschah insbesondere in der Variante, wie sie Bloom, Hastings und Madaus (1971) formuliert hatten, einfach weil dieser Text bei uns allgemeine Verbreitung fand. Die extremen und tatsächlich oft chaotischen Diskussionen, die in den USA um die Unterscheidung geführt wurden, haben im deutschsprachigen Raum allerdings nicht stattgefunden. Ein wesentlicher Grund war wohl, dass hier erst noch die Entwicklung angemessener Entwicklungs- und Schulleistungstests im Vordergrund stand.

Wie dem auch sei – jedenfalls wurde bei uns der Gedanke der formativen Evaluation und damit die Idee der systematisch wiederholten Messung einer zu erlernenden Kompetenz nicht in größerem Ausmaß aufgegriffen – und zwar auch nicht in der deutschen Sonderpädagogik, die das amerikanische „Curriculum Based Measurement“ lange Zeit nicht wahrgenommen hatte. Das änderte sich allerdings deutlich mit einem Beitrag in der Zeitschrift „Heilpädagogische Forschung“, in dem die entsprechende Entwicklung, die in den USA stattgefunden hatte, detailliert vorgestellt und positiv kommentiert worden war (Klauer, 2006). So haben dann Diehl und Hartke (2007) ebenfalls einen informativen Beitrag zur Thematik publiziert und Walter (2008) veröffentlichte eine erste und besonders umfangreiche empirische Studie zur „curriculumbasierten“ Messung der Leseleistung von Sonderschulkindern, wobei er über recht hohe Retestkorrelationen zwischen den einzelnen Tests berichten konnte. Später legte er Weiterführungen dieser Untersuchungen vor, die die sehr ermutigenden Befunde umfangreich bestätigten (Walter, 2009a, b; 2010a, b). Strathmann und Klauer (2008) führten eine Pilotstudie zum Lernverlauf des Rechtschreibens in Grundschulen durch, wobei sie – wie Fuchs und Fuchs (1993) in den USA demonstriert hatten – einzelne Wörter statt ganzer Sätze diktierten. Auf diese Weise konnte der Lernverlauf für ganze Klassen wie für einzelne Kinder dargestellt werden. Die vorgelegten Ergebnisse sind in testtheoretischer Hinsicht allerdings nicht so erfreulich wie die von Walter und die von amerikanischen Autoren berichteten Befunde. Inzwischen haben Strathmann, Klauer und Greisbach (2010) eine weitere Studie zur Entwicklung der Rechtschreibung über zwanzig Schulwochen in sechs Grundschulklassen veröffentlicht, die in mancher Hinsicht ebenfalls wenig ermutigende Ergebnisse brachte. Ferner haben Wilbert und Linnemann (2011) einen differenzierten Beitrag zur *Analyse* von Tests zur Lernverlaufsdiagnostik vorgelegt, während Souvignier und Förster (2011) in ihrer umfangreichen experimentellen Untersuchung zur prozessorientierten Diagnostik der Lesekompetenz von Viertklässlern sogar den *lernförderlichen* Aspekt der Lernverlaufsdiagnostik deutlich nachweisen konnten (vgl. auch die beiden Beiträge von Souvignier und Mitarbeitern, in diesem Band).

Ein besonderes Verdienst kommt zweifellos Jürgen Walter zu. Er brachte den ersten deutschen Test dieser Art, den Test zur „Lernfortschrittsdiagnostik Lesen“ in der Reihe „Deutsche Schultests“ heraus (Walter, 2010a), der das in den USA

bewährte Grundkonzept produktiv und bedeutsam weiterentwickelte. Inzwischen liegt ein weiteres Testverfahren zur „Verlaufsdiagnostik sinnerfassenden Lesens“ von ihm vor (Walter, 2013), während Strathmann und Klauer (2010, 2012) einen Test „Lernverlaufsdiagnostik Mathematik für zweite bis vierte Klassen“ vorlegen konnten.

Abschließend hierzu sei noch kurz auf die Terminologie eingegangen. Manche der deutschen Publikationen verwendeten einfach den Begriff „Curriculum Based Measurement“ (CBM). Wegen der Problematik des Begriffs hatte ich in der Publikation von 2006 auch den Begriff „Lernfortschrittsmessung“ eingeführt, den Walter (2010a, b) aufgriff. Da die Forschung mittlerweile aber zeigte, dass keinesfalls immer von Lern*fortschritten* die Rede sein kann, scheint der Begriff der „Lernverlaufsdiagnostik“ (Strathmann & Klauer, 2010) angemessener zu sein, der inzwischen auch verbreitet Anwendung findet.

1.3 Lernverlaufsdiagnostik als Variante formativer Evaluation

1.3.1 Wiederholte Messung ein und derselben Kompetenz

Soll der Lernverlauf und damit die Entwicklung einer Kompetenz über einen längeren Zeitraum hinweg erfasst werden, etwa über ein ganzes Schuljahr hinweg, so wird kein Weg daran vorbeigehen, die fragliche Kompetenz immer wieder zu testen, um die Leistungsentwicklung zu dokumentieren. Normalerweise würde man hierzu Paralleltests einsetzen. Wollte man aber jede Woche einen solchen Test erheben, so wären 40 Paralleltests erforderlich, da es in der Regel 40 Wochen pro Schuljahr gibt. Und wollte man Tests nur alle 14 Tage erheben, so wären immer noch 20 Paralleltests vonnöten. Tatsächlich hat Walter (2010a) in seinem Lesetest 28 Paralleltest zur Verfügung gestellt, während Souvignier und Förster (2011) in ihrer experimentellen Studie zum Effekt ihrer Lernverlaufsdiagnostik Lesen 4×2 Paralleltests heranzogen. Aber in vielen anderen Bereichen ist es zumindest unüblich, so viele Paralleltests im Vorhinein zu entwickeln und auf ihre Eignung hin zu überprüfen.

Allerdings bietet sich oft die Möglichkeit, das Lehrziel und damit auch die fragliche Kompetenz durch die *Aufgabenmenge* zu definieren, zu der die Kompetenz qualifiziert. Ist dies geschehen, so kann man aus der Aufgabenmenge repräsentative Stichproben ziehen. Dieses Vorgehen ist im Einzelnen relativ komplex, aber in der Tradition lehrzielorientierter oder kriteriumsorientierter Tests differenziert belegt und erprobt (Klauer, 1987). Beispiele für dieses Vorgehen im Bereich Mathematik Grundschule findet man in dem Beitrag von Strathmann (in diesem Band). Die dort vorgestellte Prozedur der zufallsgesteuerten Generierung von Aufgabenstichproben konnte erst im Zeitalter des PC konsequent realisiert

werden, hat aber, wie unten erläutert, bemerkenswerte testtheoretische Vorteile, die in der Tradition des CBM kaum genutzt werden konnten.

Dieses letztere Vorgehen ist jedoch mit einer Bedingung verknüpft, die auf den ersten Blick befremdlich erscheint: Man muss dann etwa das ganze Schuljahr hindurch Aufgabenstichproben vorlegen, die zwar repräsentativ sind für die Kompetenz, die am Ende des Schuljahres beherrscht sein soll, wohingegen die Kompetenz aber erst im Laufe des Schuljahres erworben wird. Die Probanden erhalten dann immer wieder Tests, von denen klar ist, dass sie noch nicht alle Aufgaben lösen können, aber dass sie im Laufe des Schuljahres imstande werden, mehr und mehr der Aufgaben zu lösen. Das ist nicht nur in der Mathematik so, sondern in nahezu allen anderen Sachfächern wie in Geschichte, Erdkunde, Biologie, Physik und dergleichen mehr. Man wird also den Kindern klar und deutlich sagen müssen, dass sie noch nicht alle Aufgaben erfolgreich bearbeiten können und deshalb die zu schweren Aufgaben unbesorgt auslassen dürfen.

Möglicherweise kann diese Problematik mit spezielleren testtheoretischen Ansätzen gelöst werden. Solange dies nicht der Fall ist, sollte man in der beschrieben Weise vorgehen.

1.3.2 Homogene Testschwierigkeiten

Ein für Lernverlaufsdiagnostik entscheidender Punkt betrifft die Schwierigkeit der einzelnen Tests. Angenommen, man würde unbeabsichtigt heute einen relativ schweren und morgen einen relativ leichten Test zur gleichen Thematik erheben, so würde vermutlich eine deutliche Leistungsverbesserung zu verzeichnen sein, ohne dass sich die Leistung tatsächlich verbessert haben müsste. Eine Lernverlaufsdiagnostik muss gewährleisten, dass die einzelnen Tests, die da gegeben werden, (erstens) dasselbe erfassen und (zweitens) auch stets gleich schwer sind.

Es wäre natürlich keine Lösung, denselben Test mehrfach später wiederholt zu erheben: Er würde alleine schon wegen der Testwiederholung leichter werden, also zu besseren Ergebnissen führen, und es ist fraglich, ob er auch immer noch dasselbe messen würde. Man wird also jeweils neue Tests geben müssen, die stets dasselbe Leistungsspektrum abdecken sollen und immer gleich schwierig sein müssen. Die *Homogenität* der Testschwierigkeit ist also für jede Art von Lernverlaufsdiagnostik und Veränderungsmessung von zentraler Bedeutung. Gerade dieser Aspekt wurde bislang nicht hinreichend berücksichtigt, ja oft gar nicht thematisiert. Vielfach wurde unterstellt, die einzelnen Tests seien alle gleich schwer.

Bei dem Versuch, die Lernverlaufsdiagnostik zur Entwicklung der Rechtschreibkompetenz in der Grundschule einzusetzen, konnten Strathmann und Klauer (2008) auf einen definierten Grundwortschatz zurückgreifen, um daraus durch

einen Zufallsgenerator Stichproben von jeweils 20 Wörtern zu ziehen. Auf diese Weise sollte gewährleistet werden, dass die Diktate immer die gleiche Kompetenz erfassen. Über längere Zeit hinweg wurden nun Grundschülern solche Wortstichproben vorgelegt. Es stellte sich aber heraus, dass die Wortstichproben mal schwerer und mal leichter waren, so dass sie sich für eine Lernverlaufsdiagnostik nicht besonders gut eigneten. Zu einem entsprechenden Ergebnis kamen auch Strathmann, Klauer und Greisbach (2010) in einer weiteren und verbesserten Studie zur Rechtschreibung. Die Autoren schlossen daraus, dass auch der bei Grundschuldidaktikern eingesetzte Grundwortschatz nicht homogen genug ist, so dass selbst Zufallsstichproben von Items nicht gleich schwere Anforderungen stellen: Man könnte entweder die Anzahl der zufällig zu ziehenden Items deutlich erhöhen oder aber den Grundwortschatz in schwierigkeitshomogenere Teilmengen zerlegen und die Teilmengen bei der Stichprobenziehung in zuvor festgelegten Proportionen berücksichtigen. Dieses letztere Verfahren wurde bei der Lernverlaufsdiagnostik Mathematik von Strathmann und Klauer (2012) erfolgreich eingesetzt (siehe den Beitrag von Strathmann, in diesem Band).

Ein weiteres Problem stellt sich, wenn es darum geht, die Homogenität der Schwierigkeit aufeinander folgender Tests *nachzuweisen*. In der Praxis rechnen wir ja damit, dass die Kinder im Laufe der Zeit etwas lernen, sich also verbessern. Gibt man dann aber Tests mit stets den objektiv gleichen Schwierigkeiten, so müssen die Tests von Mal zu Mal leichter werden – eben in dem Ausmaß, in dem sich die Kinder verbessern. Der Schwierigkeitsgrad nimmt theoretisch also im Fall des Lernens kontinuierlich ab. Strathmann und Klauer (2008, 2010, 2012) haben in dieser Situation den Ausweg gewählt, immer nur zwei direkt aufeinander folgende Tests auf homogene Schwierigkeit zu testen. Dabei muss man allerdings unterstellen, dass der Lernzuwachs in dieser Zeit vergleichsweise gering ist. Tatsächlich hat sich dieses Vorgehen bewährt.

1.3.3 Änderungssensibilität

Effektive Lernverlaufsdiagnostik setzt allerdings weiterhin voraus, dass die Verfahren in der Lage sind, Kompetenz*zuwächse* wie Kompetenz*verluste* sensibel zu erfassen. Das gilt auch für Tests, die im Rahmen der „Response to Intervention“ eingesetzt werden. Veränderungsmessung spielt traditionell in anderen Bereichen der Psychologie eine große Rolle, so in der Klinischen Psychologie. Da es sich bei Erziehung und Unterricht um langfristige Interventionen handelt, sollte man annehmen, dass Veränderungsmessung in der Pädagogischen Psychologie besonders stark vertreten sei. Das ist aber keineswegs der Fall, hauptsächlich weil es noch an hierzu geeigneten Testverfahren mangelt. Lernverlaufsdiagnostik sollte jedenfalls in der Lage sein, Zuwächse in der Kompetenzentwicklung, aber auch etwaige Lernverluste zuverlässig zu erfassen, wenn das Verfahren in regelmäßigen Abständen eingesetzt wird. Es sollte also *änderungssensibel* sein, ein Aspekt,

dem im Rahmen der Veränderungsmessung zentrale Bedeutung zukommt (Bryk & Raudenbush, 1987; Fischer & Formann, 1982; Klauer, 2011; Petermann, 1978, 2010; Petermann & Hehl, 1979). Ob ein Test änderungssensibel ist, wird zweckmäßig durch ein spezielles Experiment überprüft. Hierzu bietet sich ein Zwei-Gruppen-Versuchsplan mit einer Fördergruppe und einer Kontrollgruppe ohne Förderung an mit entsprechenden Prä- und Posttests. Die Fördergruppe sollte dann einen höheren Zugewinn als die Kontrollgruppe erzielen, die die spezielle Intervention ja nicht erhalten hat. Lässt sich ein solcher differentieller Zugewinn nachweisen, so hat man gezeigt, dass (a) die Förderung wirksam war, was jetzt aber nur mittelbar interessiert, und dass (b) der Test Zuwächse der fraglichen Kompetenz auch diagnostizieren kann. Auf diese Weise haben Klauer und Strathmann (2013) die Änderungssensibilität ihrer „Lernverlaufsdiagnostik Mathematik“ nachgewiesen.

1.3.4 Welche Testtheorie kommt infrage?

Tack (1986) hatte schon vor Jahren nachgewiesen, dass klassisch konstruierte Tests praktisch nicht geeignet sind, Änderungen zu erfassen. Insbesondere lässt sich mit klassisch konstruierten Tests das Reliabilitäts-Validitäts-Dilemma nicht überwinden. Hohe Retest- oder Paralleltestreliabilitäten sind nicht vereinbar mit unterschiedlichen Veränderungen. Sie sprechen vielmehr dafür, dass die Probanden sich in ihren Werten nicht oder alle im gleichen Ausmaß und in der gleichen Richtung verändert haben. Beispiel: Haben alle Probanden beim nächsten Test zwei Punkte mehr erzielt, so resultiert eine Korrelation zwischen den beiden Test von $r = 1$. Haben die Probanden aber unterschiedlich zugelegt, manche sich verschlechtert und andere sich verbessert, so müssen niedrigere Retestreliabilitäten resultieren. Änderungssensible Verfahren sollten daher nur *mäßige* Retest- oder Paralleltestreliabilitäten aufweisen, aber hohe Werte der Split-half Reliabilität.

Es stellen sich Probleme, will man auf dem Boden der klassischen Testtheorie änderungssensible Tests entwickeln. Itemschwierigkeiten und Itemtrennschärfen spielen dabei bekanntlich zentrale Rollen. Wenn nun aber immer *neue* Aufgaben gegeben werden, so ist es schlicht sinnlos, Itemschwierigkeiten und Itemtrennschärfen ermitteln zu wollen. Fehlen aber diese Parameter, so ist es nicht möglich, klassische Paralleltests zu entwickeln. Die Retestreliabilität lässt sich dann nicht einmal bestimmen, da ja kein Test wiederholt wird, denn es sind immer andere Tests, die gegeben werden. Ebenso wenig kommt die Bestimmung der Inneren Konsistenz mittels Cronbachs α infrage: Dann müsste nämlich eine Stichprobe von Probanden dieselben Testaufgaben erhalten, damit eine Personen-Item-Matrix erstellbar wird. Wenn aber jeder Teilnehmer eine eigene Teststichprobe erhält, lässt sich auch Cronbachs α nicht berechnen. Kurz und gut: Änderungssensible Verfahren gemäß der klassischen Testtheorie zu konstruieren ist praktisch ausgeschlossen.

In der Item Response-Theorie (IRT) gibt es interessante Modelle, von denen das eine oder andere für die Lernverlaufsdiagnostik relevant werden könnte. Das gilt insbesondere für Prozessmodelle, wie sie schon Spada und Kempf (1977) vorgestellt hatten und wie sie Scheiblechner (1996) zusammenfassend beschreibt. Immerhin ist es nicht ungewöhnlich, bei zwei Tests, die Vergleichbares erfassen sollen zu prüfen, ob sie Rasch-homogen sind. Ist dies der Fall, so kann man sie auf eine gemeinsame Fähigkeitsdimension normieren. Theoretisch ist sogar denkbar, dass eine hinreichend große Anzahl von Items Rasch-homogen ist, so dass man des Öfteren immer neue Stichproben daraus ziehen könnte. In einem solchen Fall könnte man mit dem Rasch-Modell oder, je nach den Bedingungen, mit einem anderen Latent Trait-Modell Lernverlaufsdiagnostik durchführen. Auf weitere hier interessante Möglichkeiten sind Rost und Spada (1983) schon vor Jahren eingegangen. Wilbert und Linnemann (2011) verwendeten jedenfalls schon ein probabilistisches Testmodell im Rahmen der Lernverlaufsdiagnostik, ebenso wie Strathmann und Klauer (2012), deren „Lernverlaufstest Mathematik“ auf dem Binomialmodell beruht.

1.3.5 Typische Verlaufskurven

Hat man ein Verfahren zur Verfügung, um schwierigkeitshomogene und änderungssensible Tests zu erzeugen, so führt deren Anwendung zu einer Vielfalt von Verläufen. Die beiden Verläufe, die hier nur zur ersten Illustration geboten werden, beziehen sich auf Mathematiktests einer vierten Klasse, wobei über das ganze Schuljahr hinweg jede zweite Woche ein Test durchgeführt wurde, wobei die Tests sich auf die am Ende des Schuljahres zu erwartende Kompetenz bezogen. Die Lernverläufe basieren also auf 20 Testerhebungen, bei denen jeweils 24 Aufgaben gestellt worden waren (zu den Einzelheiten des Tests und der Datenerhebung siehe den Beitrag von Strathmann, in diesem Band).

Die Lernverläufe der einzelnen Probanden unterscheiden sich oft deutlich von den Verläufen, die auf den Mittelwerten der ganzen Klasse beruhen. Bei diesen letzteren begegnet man vergleichsweise oft *linearen Anstiegen*, weil etwa das ganze Schuljahr hindurch mehr oder minder stetig anwachsende Leistungen zu beobachten sind. Klassen unterscheiden sich erfahrungsgemäß (a) in dem Niveau, mit dem sie das Schuljahr beginnen, und (b) in der Steigung, also im Anstieg. Manche Klassen fangen sozusagen am Nullpunkt an, während andere schon beim Schuljahrsbeginn einen beachtlichen Teil des Ziels erreicht haben, um das es in dem Schuljahr geht. In solchen Fällen stoßen die Verlaufskurven oft schon relativ früh an die „Decke“, wenn nämlich das Lehrziel vorzeitig erreicht wird. Dann hat man es mit einem Ceilingeffekt zu tun, der sich auch als nichtlineare Entwicklung darstellt, denn es resultieren am Ende negative Steigungen. Ist der Anstieg weniger steil, so hat man es unter Umständen zwar die ganze Zeit über mit linearen Verläufen zu tun, die aber das Lehrziel am Ende verfehlen wie dies am Beispiel der Abbildung 1 deutlich wird.