Empirische Sonderpädagogik, 2011, Nr. 3, S. 207-224

Lernverlaufsdiagnostik – Konzept, Schwierigkeiten und Möglichkeiten

Karl Josef Klauer

RWTH Aachen

Ähnlich wie bei der curriculumbasierten Messung (CBM) werden bei der Lernverlaufsdiagnostik über einen längeren Zeitraum hinweg regelmäßig Tests gegeben, die stets ein und dieselbe Kompetenz messen sollen und dann erlauben, den Lernverlauf abzubilden. Solche Tests müssen immer neu sein, jedes Mal aber dasselbe messen und stets gleich schwierig sein, was unerwartet problematisch ist. Darüber hinaus sollen sie in der Lage sein, Änderungen der Kompetenz der Probanden sensibel zu diagnostizieren. Klassisch konstruierte Tests dürften für die Lernverlaufsdiagnostik schwerlich in Frage kommen können. In diesem Beitrag wird ein Lösungsvorschlag für die meisten der Probleme offeriert. Der Vorschlag beruht auf der Grundlage lehrzielorientierter Tests und setzt auf Itemsampling in Verbindung mit dem binomialen Testmodell. Abschließend wird ein auf diese Weise konstruiertes Verfahren zur Lernverlaufsdiagnostik im Bereich Mathematik Grundschule vorgestellt, das demnächst verfügbar sein wird.

Schlüsselwörter: Lernverlaufsdiagnostik, Itemsampling, Veränderungsmessung, lehrzielorientierte Tests, binomiales Testmodell, Mathematiktests

Diagnosing the Course of Learning – Concept, Difficulties and Chances

To diagnose the course of learning during a longer period of time as it is similarly practiced in curriculum-based measurement (CBM) one has to use tests which always are equally difficult and every time measure the same competence. Until today it is not clear how to construct a greater number of tests of equal difficulty und validity. Moreover, it is not clear which test theory is suitable for such tests since classical test theory turns out to be problematic with respect to such tests. Finally, suitable tests should be sensitive even to minor changes of competence. In this article a solution is offered based upon criterion-referenced tests, item sampling and the binomial test model. Finally, a forthcoming computer-based test is described which will overcome most of the problems and which can be used to measure the course of learning of mathematics with children of elementary school.

Key words: curriculum-based measurement, item sampling, measurement of change, criterion-referenced tests, binomial test model, mathematics tests

Bei der Lernverlaufsdiagnostik geht es darum, statt einer punktuellen Testerhebung regelmäßig in relativ kurzfristigen Abständen die Entwicklung einer Kompetenz über längere Zeit hinweg durch geeignete Tests zu verfolgen. Dabei soll es sich um immer neue Tests handeln, die aber stets die Ausprägung ein und derselben Kompetenz messen. Diese Art der Diagnostik wurde bekanntlich an der Universität Minnesota in Minneapolis von

Stanley Deno und einigen seiner Doktoranden in sonderpädagogischem Kontext seit 1972 entwickelt. In Deutschland ist das Verfahren jahrzehntelang überhaupt nicht beachtet worden, auch nicht in der Sonderpädagogik. Erst ein Artikel von mir (Klauer, 2006) über diese Thematik in einer sonderpädagogischen Zeitschrift führte zu beachtlichen Reaktionen. So haben Diehl und Hartke (2007) einen informativen Beitrag publiziert und Walter (2008) veröffentlichte eine besonders umfangreiche empirische Studie zum Lernverlauf beim Lesen von Sonderschulkindern, wobei er über recht hohe Retestkorrelationen berichten konnte. Später legte er Weiterführungen dieser Untersuchungen vor, die die sehr ermutigenden Befunde umfangreich bestätigten (Walter, 2009a, 2009b; 2010a; 2010b). Strathmann und Klauer (2008) führten eine Pilotstudie zur Entwicklung des Rechtschreibens durch, wobei sie, wie dies Fuchs und Fuchs (1993) in den USA demonstriert hatten, einzelne Wörter statt ganzer Sätze diktierten. Die vorgelegten Ergebnisse sind 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 in sechs Grundschulklassen veröffentlicht, die in mancher Hinsicht ebenfalls ernüchternde Ergebnisse brachte. Dies hängt möglicherweise zusammen mit einem speziellen Problem, auf das noch einzugehen sein wird.

Ein besonderes Verdienst kommt zweifellos Jürgen Walter zu. Im gleichen Jahr hat Walter (2010a) den ersten deutschen Test dieser Art, den Test zur „Lernfortschrittsdiagnostik Lesen" in der renommierten Reihe „Deutsche Schultests" vorgelegt, der das in den USA bewährte Grundkonzept produktiv und bedeutsam weiterentwickelte.

Zur Terminologie

Es erscheint aber angemessen, zunächst auf die Terminologie einzugehen. Für diese Methoden hat Deno die Bezeichnung *„curriculum-based measurement"* (CBM) eingeführt. Mit *„curriculum-based"* war gemeint, dass die Beherrschung der im aktuellen Unterricht tatsächlich vermittelten Inhalte erfasst werden sollte. Insofern hebt sich das Verfahren deutlich von klassischen Schulleistungstests ab, die überregional standardisiert sind und sich an den über Jahre geltenden Lehrplänen orientieren. Stattdessen sollte die CBM verfolgen, was hier und jetzt gelehrt und gelernt worden ist. Tatsächlich wurden diese Beschränkungen in der Praxis etwa der Lesetests bald überholt, so dass die Bezeichnung kaum mehr diesen Aspekt der CBM repräsentiert.

Als sinngemäßer statt wörtlicher Eindeutschung hatte ich den Begriff der *Lernfortschrittsmessung* für das CBM-Verfahren gewählt, einfach weil damit das zentrale Anliegen des Verfahrens besser gekennzeichnet würde (Klauer, 2006). Analog hat Walter seinen Lesetest als „Lernfortschrittsdiagnostik Lesen" bezeichnet. Tatsächlich geht es ja darum, die Lernentwicklung und den Lernfortschritt einzelner Kinder wie ganzer Klassen über längere Zeit hinweg zu dokumentieren. Erste umfangreiche Erprobungen im Bereich Rechtschreibung, Mathematik und beim Lesenlernen (Strathmann & Klauer, 2008; Strathmann, Klauer & Greisbach, 2010; Strathmann & Klauer, 2010; Grosche & Hintz, 2010) zeigten allerdings deutlich, dass von Lern*fortschritt* keineswegs immer die Rede sein kann: Tatsächlich gibt es nicht nur Lern*stillstände*, sondern auch Lern*verluste*, mehrfache Richtungswechsel des Verlaufs und asymptotisch verlaufende Lernkurven, die am Ende keine Zuwächse mehr zeigen, was zu einer Fülle verschiedenartigster Lernverläufe führt. Diese Beobachtungen lassen es geraten erscheinen, das Vorhaben allgemeiner als *Lernverlaufsdiagnostik* zu kenn-

zeichnen. Tatsächlich gibt es Verlaufsdiagnostik auch in anderen Bereichen wie etwa in der Therapie, so dass es sich um einen bekannten Terminus handelt, der nur in der pädagogisch-psychologischen Diagnostik bislang kaum relevant war.

Da es sich aber trotz der bedeutenden Vorarbeiten in den USA noch um eine relativ neue Art der Diagnostik in unserem Feld handelt, wird man sich nicht wundern, auf eine Reihe ungelöster Probleme zu stoßen. Auf einige wichtige sei daher zunächst eingegangen.

Probleme der Lernverlaufsdiagnostik

Schwierigkeit der Tests

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 *realiter* 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. Schwankt aber deren Schwierigkeitsniveau, so wird man unvermeidbar Schlussfolgerungen ziehen, die in keiner Weise sachlich gerechtfertigt sind. 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. In aller Regel wurde unterstellt, die einzelnen Tests seien alle gleich schwierig.

Bei einem 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 das gleiche Leistungsspektrum abdecken. Über längere Zeit hinweg wurden nun Grundschülern solche Wortstichproben vorgelegt, eben analog zu dem Vorgehen amerikanischer Kollegen beim CBM. 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 schließen 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 – was dem Grundanliegen der Lernverlaufsdiagnostik widerspräche –, oder aber den Grundwortschatz in schwierigkeitshomogenere Teilmengen zerlegen und danach ein angemesseneres Verfahren der Stichprobenziehung anwenden.

So muss man also damit rechnen, bei der Lernverlaufsdiagnostik auf Probleme zu stoßen, wenn es darum geht, Tests zu erzeugen, die immer gleich schwer sein sollen. Bei seinem Test „Lernfortschrittsdiagnostik Lesen" hat Walter (2010) beispielsweise den Flesch-Index eingesetzt, der die Lesbarkeit von Tex-

ten einschätzen lässt. Es gibt mehrere solcher Flesch-Indizes; der erste stammt von dem Österreicher Rudolf Flesch selbst, der 1938 in die USA auswanderte und dort sehr erfolgreich seinen Index propagierte. Der Index hängt von der durchschnittlichen Länge der Wörter und Sätze ab, unterstellt also, dass kurze Wörter und kurze Sätze leichter lesbar sind. Die von Walter ausgewählten 28 Leseabschnitte gehören gemäß Flesch-Index zu den leicht lesbaren Texten. Zwischen der bei Schulkindern erhobenen Leseleistung und dem Flesch-Index besteht aber kein praktisch bedeutsamer Zusammenhang, wie dies auch Walter (2010a, S. 15) zeigen konnte.

Ein weiteres Problem stellt sich, wenn es darum geht, die Schwierigkeit der aufeinander folgenden Tests zu *messen*. 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) 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 vernachlässigbar gering ist.

Für die Grundschulmathematik haben Strathmann und Klauer (in Vorbereitung) ein Verfahren gewählt, das es gestattet, Tests zu erzeugen, die jeweils dasselbe Konstrukt messen und sich auch noch durch homogene Schwierigkeiten auszeichnen. Darauf wird weiter unten kurz einzugehen sein.

Validität der Tests

Sollen mehrere Tests alle die Ausprägung ein und derselben Kompetenz erfassen, also wirklich dasselbe messen, so müssen sie die gleiche Validität aufweisen.

Speziell der CBM ging es außerdem ursprünglich noch darum, genau das zu erfassen, was in eben dieser Klasse unterrichtet wurde. Vom Konzept her sollte ja gerade hier ein entscheidender Vorteil gegenüber standardisierten Tests liegen, die natürlich nicht darauf eingehen können, was hier und jetzt gelehrt und gelernt wird. Um genau das zu testen, was unterrichtet wurde, bietet sich allerdings ein Ausweg an, der in den USA außerhalb der CBM-Tradition wohl öfter eingesetzt wird, indem man nämlich schlicht auf das hin unterrichtet, was der standardisierte Test fordert. Der Ausweg stellt aber das vernünftige Verfahren auf den Kopf.

Üblicherweise geht man in der CBM-Tradition so vor, dass vom gleichen Lehrstoff mehr oder minder zufällig Stichproben gezogen und als Test präsentiert werden. Beim Lesen etwa schlägt man ein geeignetes Buch auf und lässt die einzelnen Schüler an beliebigen Stellen für eine Minute laut lesen, um die Lesefehler auszuzählen. Walter (2010a) hat diese Methode aus guten Gründen für seinen Lesetest nicht übernommen, denn wer garantiert, dass die Textstellen immer die gleichen Anforderungen stellen? Ähnlich wird auch in der Mathematik vorgegangen, dass man etwa stets eine kleine Anzahl von Aufgaben der Art stellt, die eben behandelt werden, und ermittelt, wie gut das einzelne Kind die Aufgaben bewältigt. Aber auch innerhalb eines Schuljahres steigen die Anforderungen, die in der Mathematik gestellt werden, etwa wenn zunächst innerhalb der Zehner addiert wird, um danach auch Aufgaben zu bringen, die die Überschreitung der Zehner erfordern. Dabei wird das Lehrziel klar geändert und entsprechend die Validität solcher Testprozeduren. Wie soll man aber gar im Sachunterricht, in der Erdkunde, Geschichte oder der Biologie über längere Zeit hinweg Tests geben, die stets das Gleiche messen, also von gleicher Validität sind?

Eine Lernverlaufsdiagnostik, die über einen längeren Zeitraum von etwa einem ganzen Schuljahr durchgeführt wird, muss sich von der Forderung lösen, genau das zu messen, was hier und jetzt im Unterricht gelehrt und gelernt wird. Vielmehr kann es nur darum gehen, das zu messen, was die Kinder am Ende des Schuljahres können sollen. Nur in diesem Fall wird man zeigen können, wie sich die Leistung der Kinder im Laufe der Zeit mehr oder minder zügig dem Kompetenzniveau annähert, das gefordert ist. Hierzu gibt es in der CBM-Tradition meines Wissens noch keine brauchbare Lösung.

Um welche Art von Tests handelt es sich eigentlich?

Vergleichsweise wenig beachtet wurde außerdem die Frage, auf welche Testtheorie man sich in der pädagogisch-psychologischen Verlaufsdiagnostik berufen kann. In diesem Abschnitt wird kurz gezeigt werden, dass Tests, die die genannten Bedingungen erfüllen, schwerlich gemäß der klassischen Testtheorie konstruierbar sind. Allerdings gibt es inzwischen alternative testtheoretische Ansätze, doch gewinnen sie trotz ihrer unbestreitbaren Vorteile in der derzeitigen diagnostischen Praxis nicht sehr an Boden. Es wird dennoch zu prüfen sein, ob man bei diesen Theorien und Modellen fündig werden kann für die Konzeption einer Lernverlaufsdiagnostik im vorliegenden Sinne.

Wieso ist es problematisch, Tests der hier geforderten Art auf dem Boden der *klassischen Testtheorie* zu entwickeln? Das soll nun an einigen Beispielen kurz erläutert werden. Itemschwierigkeiten und Itemtrennschärfen spielen bei der Testkonstruktion gemäß der klassischen Testtheorie zentrale Rollen. Für die Auswahl geeigneter Aufgaben sind beide Parameter heranzuziehen, wobei die Schwierigkeiten möglichst über einen weiten Bereich streuen und die Trennschärfeindizes so groß wie möglich ausfallen sollen. Analoges gilt auch für die Entwicklung von Paralleltests, die das gleiche Konstrukt erfassen sollen. Man wählt dazu Items, die in beiden Parametern nahezu gleiche Werte aufweisen, also nahezu gleich schwer und gleich trennscharf sind, um dann per Zufall die Items dem ersten oder dem zweiten Test zuzuweisen. Paralleltests dieser Art kommt eine wichtige Bedeutung zu in der Ermittlung der Reliabilität des Tests, also einem zentralen Anliegen der klassischen Testtheorie.

Wenn nun aber immer andere Aufgaben gegeben werden, so ist es schlicht sinnlos, Itemschwierigkeiten und Itemtrennschärfen ermitteln zu wollen, was ja auch keiner in der CBM-Tradition tut. Von Paralleltests kann also auch keine Rede sein. Die Retestreliabilität lässt sich ebenfalls nicht bestimmen, da ja kein Test wiederholt wird, denn es sind immer neue Tests, die gegeben werden. Ebenso wenig kommt die Bestimmung der Inneren Konsistenz mittels Cronbachs α in Frage: 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: Die Verfahren der Lernverlaufsdiagnostik gemäß der klassischen Testtheorie zu konstruieren ist praktisch ausgeschlossen.

Wie sieht es aber mit den *Latent-Trait-Modellen* und speziell auch mit deren probabilistischen Varianten aus, wie Roskam (1996) sie übersichtlich dargestellt hat? Für alle Modelle dieser letzteren Art und demnach auch für Rasch-Modelle gilt, dass sie Annahmen über die Itemcharakteristiken voraussetzen. Danach ist die Wahrscheinlichkeit, dass ein Item gelöst wird, eine Funktion des Fähigkeitsparameters der Person und des Itemparameters. Im Rasch-Modell muss es sich sogar um eine logistische Funktion des Itemparameters handeln. Wenn aber immer neue Items gegeben werden, so lässt sich zunächst auch kein solches Modell anwenden. Allerdings ist das zweifellos nicht das letzte Wort in dieser Sa-

che. 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 Kempf (1977) vorgestellt hatte 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.

Es dürfte also angezeigt sein, in diesem theoretischen Umfeld nach bislang nicht beachteten Lösungen zur Lernverlaufsdiagnostik zu suchen.

Änderungssensibilität

Prinzipiell sollte der Veränderungsmessung im Rahmen der pädagogisch-psychologischen Diagnostik ein größeres Gewicht zukommen, doch ist das bislang kaum der Fall, wohingegen die Thematik in Therapieverlaufsstudien schon seit Längerem starke Beachtung findet (vgl. Petermann, 1978; 2010; Petermann & Hehl, 1979). Bei der Itemanalyse zur Konstruktion von Tests gibt es zwar schon seit Längerem die Möglichkeit, änderungssensible Items nach einer zweiten Testung zu ermitteln. Die Möglichkeiten hierzu hat Krauth (1995) in seinem Lehrbuch an verschiedenen Stellen ausführlich dargelegt. Allerdings geht es hier nicht nur um die Auswahl änderungssensibler Items, denn die gesamte Folge der Tests soll ja änderungssensibel sein.

Es bleibt aber das Verdienst von Jürgen Walter, in der deutschsprachigen Literatur zur Lernverlaufsdiagnostik als erster dem Aspekt der Änderungssensibilität die notwendige Bedeutung zuerkannt zu haben. Soweit ich sehe, wurde die Thematik der Änderungssensibilität auch in der CBM-Literatur nicht so hinreichend wie wünschenswert untersucht. Wenn es aber darum geht, Leistungsfortschritte zu diagnostizieren, genauer *Veränderungen* von Fähigkeiten oder Kompetenzen, so sollten Instrumente hierfür entwickelt werden, die nachweislich in der Lage sind, solche Veränderungen sensibel zu erfassen.

Die hierzu übliche und vielfach auch angewandte Versuchsanordnung besteht darin, dass der Leistungsstand erhoben wird, danach ein Treatment erfolgt, um anschließend den Leistungsstand erneut zu testen. Dabei ist es zweckmäßig, eine Kontrollgruppe ohne Treatment einzusetzen, alleine schon um den Einfluss der Testwiederholung berücksichtigen zu können. Es handelt sich also um einen klassischen Zwei-Gruppen-Versuchsplan mit Prätest und Posttest, wobei nur der Versuchsgruppe, nicht aber der Kontrollgruppe ein Treatment geboten wird. Das Treatment würde hier aus einem bewährten Förderunterricht bestehen, der die Lese- oder Rechenfertigkeit, eben die Fähigkeit fördert, um die es gerade geht. Der Förderunterricht sollte nachweislich wirksam sein, aber nicht allzu lange dauern, denn es soll ja nachgewiesen werden, dass das Testverfahren auch schon kleinere Änderungen zuverlässig erfassen kann.

Für die Lernverlaufsdiagnostik stehen solche experimentellen Nachweise noch aus. Walter (2010a) hat zwar durch umfangreiche Studien zeigen können, dass seine „Lernfortschrittsdiagnostik Lesen“ zu durchgängig besseren Mittelwerten führt, wenn man den Test von Klasse 1 bis Klasse 9 in Grund- und Hauptschulen anwendet, aber auch, wenn er in der Förderschule in den Klassen 5, 7 und 8

eingesetzt wird. Würde man einen herkömmlichen Lesetest für die 2. Klasse in zunehmend höheren Klassen einsetzen, so würde man sicher ebenfalls auf immer höhere Mittelwerte stoßen. Vielleicht ist diese Art des Vorgehens dann doch zu unspezifisch, um die Änderungssensibilität zu belegen.

Ein Lösungsansatz

Im Folgenden wird ein testtheoretisches Konzept vorgestellt, das für die wichtigsten der bestehenden Schwierigkeiten eine Lösung bietet. Das gilt insbesondere für die Validität des Verfahrens sowie für die Frage nach einem spezifisch geeigneten Testmodell. Dieser Ansatz eröffnet weiterhin gute Voraussetzungen für die Reliabilität der Tests, für Verlaufsanalysen, für das Problem der Homogenität der Testschwierigkeiten sowie für die Testung der Änderungssensibilität. Er basiert auf dem Konzept der kriteriumsorientierten oder lehrzielorientierten Tests. Nähere Einzelheiten zu diesem Thema findet man in einem entsprechenden Lehrbuch (Klauer, 1987).

Kriteriumsorientierte beziehungsweise lehrzielorientierte Tests

Beim schulisch relevanten Lehren geht es stets darum, spezielles Wissen oder spezielles Können zu vermitteln und zu prüfen, ob die Kinder das gelernt haben, was sie lernen sollten. So geht es auch hier darum, latente Variablen oder latente Dimensionen zu messen, also Größen, die nicht direkt erfassbar sind. Es handelt sich in allen Fällen um Kompetenzen, wobei Lehrziele beschreiben, welche Kompetenzen es im Einzelnen sind, die erreicht werden sollen.

Das Konzept lehrzielorientierter Tests lässt sich im Kern kurz umschreiben. Entscheidend ist die Forderung, die jeweilige Kompetenz, um die es im Einzelnen geht, durch eine Menge von Aufgaben zu definieren, zu deren Lösung die Kompetenz qualifiziert. Die Aufgabenmengen, deren Lösung beherrscht werden soll, sind streng im mengentheoretischen Sinne so zu definieren, dass für jede beliebige Aufgabe eindeutig entscheidbar ist, ob sie Element der Menge ist oder nicht. Solche Aufgabenmengen können einerseits dazu dienen, den Unterricht zielgerichtet zu gestalten, und andererseits dazu, den Lernerfolg zu testen. Lehrziel- oder Kontentvalidität ist – streng genommen – nur dann gegeben, wenn die Testaufgaben die definierte Grundmenge repräsentieren, wenn sie also repräsentative Stichproben der Grundmenge darstellen.

Nun gibt es mindestens drei Möglichkeiten, Mengen zu definieren – und sie sind alle drei pädagogisch relevant.

Die einfachste Definitionsvariante besteht darin, alle zur Menge gehörigen Aufgaben aufzulisten. Das gilt zum Beispiel für die Hauptstädte aller Länder oder für die Aufgaben des Einmaleins. Setzt man beispielsweise 3 x 4 = 4 x 3, so reduziert sich die Anzahl der hundert Einmaleinsaufgaben, die die Kinder lernen müssen, auf 45 verschiedene Aufgaben. Und nimmt man noch die Reihe mit 1 heraus (1 x 1, 2 x 1 usw.), so hat man es nur noch mit 35 verschiedenen Aufgaben zu tun, die entsprechend intensiv geübt werden müssen. Ein anderes Beispiel bietet der Lesetest von Walter (2010a), der genau 28 sorgfältig ausgewählte Textabschnitte bietet, die zu lesen sind und jeweils ein Urteil über die Lesefertigkeit ermöglichen.

Die zweite Variante besteht darin, mit einem Sachverhalt zu beginnen, um den es geht, und ihn im ersten Schritt durch eine Menge von Aussagensätzen komplett darzustellen. Die Aussagensätze können dann im zweiten Schritt nach bewährten Verfahren (Klauer, 1987, S. 33 ff) in eine Menge von Testaufgaben umgewandelt werden. Dabei handelt es sich um die Grundmenge von Aufgaben, die es zu lehren und zu lernen gibt. Dieses Verfahren eignet sich für Lehrstoffe

des Sachunterrichts, etwa der Biologie, der Erdkunde oder Geschichte.

Schließlich gibt es als dritte die Möglichkeit, eine Aufgabenmenge durch eine Aussagenform zu definieren. Eine Aussagenform ist eine Aussage, die mindestens eine Variable statt einer Konstanten enthält. Die Aussagenform a + b = c (a, b, c € N) enthält drei Variablen und definiert die Gesamtheit aller Additionsaufgaben in der Menge der natürlichen Zahlen und ist unendlich groß. Man kann sie natürlich in vieler Hinsicht begrenzen, so etwa N auf die Menge der natürlichen Zahlen bis 1000 beschränken, was etwa für das zweite Schuljahr relevant sein dürfte.

Da wir es in dem unten zu erläuternden Beispiel mit (Grundschul-)Mathematik zu tun haben, liegt es nahe, hier auf diese dritte Form der Definition von Mengen näher einzugehen. Auch in der Grundschulmathematik geben die Definitionen der Aufgabenmengen inhaltlich an, welche Lehrziele es sind, die erreicht werden sollen. Hier ein Beispiel, das das Vorgehen näher erläutern soll. Es handelt sich um die Definition der Subtraktionsaufgaben, die im zweiten Schuljahr gelehrt und gelernt werden sollen (E = Einer, Z = Zehner).

Die Tabelle definiert für das zweite Schuljahr die Kompetenz der Subtraktion. Faktisch wird die Aufgabenmenge durch sechs Teilmengen definiert, die die Kinder am Ende gemäß Lehrziel beherrschen sollen. Tabelle 1 legt also die Grundgesamtheit der Aufgaben fest, um die es geht. Für lehrzielorientierte Tests und entsprechend auch für die Lernverlaufsdiagnostik sind nun repräsentative Stichproben aus der Grundgesamtheit zu ziehen, so dass jeder einzelne Test die Grundgesamtheit valide abbildet. Dabei ist im Einzelnen noch konkret festzulegen, wie stark die Teilmengen in einer repräsentativen Aufgabenstichprobe vertreten sein sollen. In Tabelle 1 ist das durch die Angabe geschehen, dass aus den sechs Teilmengen je eine Aufgabe gezogen werden soll. Mit dieser Festlegung ist die Kompetenz klar definiert, deren Entwicklung über die Zeit hinweg verfolgt werden soll. Ein solches Vorgehen gewährleistet erst die Content- oder Lehrzielvalidität. Es ist, soweit ich sehe, im Kontext der curriculumbasierten Messung nie eingesetzt worden.

Repräsentative Itemstichproben können auf zweierlei Weise hergestellt werden, wenn die Grundgesamtheiten entsprechend definiert worden sind. Sind die Grundmengen in sich relativ homogen strukturiert, so kann man mittels eines Zufallsgenerators direkt Zufallsstichproben von Aufgaben ziehen oder generieren lassen. Setzen sich die Grundmengen wie in Tabelle 1 aus klar unterscheidbaren und in sich prinzipiell homogenen Teilmengen zusammen, die unterschiedliche Anforderungen stellen, so wählt man zweckmäßig ein proportional-zufälliges Verfahren der Aufgabenerzeugung. In dem Fall wird zuvor

Tab. 1: Beispiel zur Definition der Subtraktion zweites Schuljahr

	c - b =	**c - ___ = a**
Z - Z ≥ 0	1	
ZE - Z	1	
ZE - E = E	1	
ZE - E = ZE**	1	
ZE - ZE *		1
ZE - ZE**		1

* Erstes E > zweites E **Erstes E < zweites E (Zehner nach unten überschreiten)

festgelegt, zu welchem Anteil die Teilmengen in den Itemstichproben vertreten sein sollen, um danach die Aufgaben per Zufallsgenerator erzeugen zu lassen. Hierzu braucht man natürlich einen PC, auf dem einerseits die Definitionen und andrerseits ein Zufallsgenerator installiert sind – etwa mittels einer CD. Das ist heute aber kein unüberwindliches Problem mehr, wie noch deutlich werden wird.

Auf diese Weise erzeugte Zufallsstichproben von Aufgaben sind kontentvalide. Kontentvalidität der Tests ist dann durch die festgelegte Herstellungsprozedur gewährleistet.

Itemsampling und Binomiales Testmodell

Im Zusammenhang mit Testverfahren spricht man von Itemsampling, wenn *jeder* Proband eine *eigene* Zufallsstichprobe von Testaufgaben erhält. Dabei geht es also nicht darum, ein einziges Mal eine Zufallsstichprobe von Aufgaben zu erzeugen – etwa um einen Test für alle in Frage kommenden Probanden zu entwickeln, vielmehr erhält jeder Proband seine eigene Zufallsstichprobe von Aufgaben. Wenn also die Grundmenge auf einem Computer installiert ist und ein Zufallsgenerator, der nach festgelegtem Verfahren Aufgabenmengen zufällig erzeugt, so kann jede Lehrkraft beliebig viele Aufgabenblätter ausdrucken lassen, die alle dieselbe Leistung erfordern, aber mit stets anders zusammengesetzten Aufgaben (s. Strathmann & Klauer, 2010; Strathmann & Klauer, in Vorbereitung).

Das Itemsampling auf der Basis präzise definierter Grundmengen hat darüber hinaus einen entscheidenden testtheoretischen Vorteil. Lord und Novick haben bereits 1968 (S. 234 ff; S. 523 f) hergeleitet, dass man es mit einem *Binomialmodell* zu tun hat, wenn jeder Proband eine eigene Zufallsstichprobe von Aufgaben erhält. Gross und Shulman (1980) haben später eine kürzere Herleitung des Zusammenhangs vorgelegt, dass nämlich die Probandenscores binomial verteilt sind, wenn solche Itemstichproben eingesetzt werden. Nähere Einzelheiten findet man auch andernorts (Klauer, 1987, S.137 f). Hier mögen folgende Hinweise genügen.

Beim Binomialmodell wird angenommen, dass ein Proband den Fähigkeitsparameter p besitzt ($0 \leq p \leq 1$), Aufgaben der Grundmenge zu lösen. Ferner setzt das Modell voraus, dass die einzelnen Aufgaben mit 1 oder 0 bewertet, also gelöst oder nicht gelöst werden. Das Binomialmodell gilt ferner nur unter zwei Voraussetzungen: Entweder sind alle Testaufgaben gleich schwer oder jeder Proband bekommt eine eigene Zufallsstichprobe von Aufgaben, also im Fall des Itemsamplings. Beim Itemsampling können die Aufgaben unterschiedlich schwer sein. In beiden Fällen ist zu erwarten, dass der Proband x von n Aufgaben richtig löst gemäß seiner Fähigkeit p:

$$x = np$$

Die Varianz des Schätzwertes x stellt sich dann dar als

$$s^2 = np(1-p).$$

Dabei wird deutlich, dass die Varianz am größten ist, wenn $p = 0.5$ ist und dass sie deutlich abnimmt, wenn der Fähigkeitsparameter p in die Nähe von 1 oder 0 rückt. Beispiel: Gibt man 20 Aufgaben bei $p = 0.5$, so resultiert eine Varianz von $20 \times 0.5^2 = 5$. Bei $p = 0.9$ oder bei $p = 0.1$, also bei einem fast perfekten Könner oder Nichtkönner resultiert eine Fehlervarianz von $20 \times 0.9(1 - 0.9) = 20 \times 0.09 = 1.8$. Die Messgenauigkeit wird also in den Extrembereichen deutlich besser. Bei fast perfekten Könnern resultiert eine rechtsschräge Verteilung, bei fast perfekten Nichtkönnern eine linksschräge Verteilung und bei einem Fähigkeitsparameter von $p = 0.5$ hat man es mit einer symmetrischen Verteilung zu tun. Die Binomialverteilung strebt gegen die Normalverteilung, wenn das n, die Anzahl der Aufgaben, hinreichend groß wird.

Das Testmodell ist also keineswegs neu. Praktisch umgesetzt wurde es bislang nur wenig und dann im Rahmen der kriteriumsorientierten Tests. Dabei wurden zumeist gleich schwere Testaufgaben konstruiert, einfach weil das Itemsampling früher kaum zu realisieren war. Das ist heute nun anders. Denn erst in jüngerer Zeit stehen Lehrkräften PCs zur Verfügung, die hier praktikable Lösungen zulassen. Zieht man also kontentvalide Zufallsstichproben heran, von denen jeder Teilnehmer eine eigene erhält, so hat man es mit einem bekannten Testmodell zu tun, und der zu schätzende *Personenparameter* p stellt sich einfach dar als der Anteil (oder Prozentsatz) richtig gelöster Aufgaben:

$$p = \frac{x}{n}$$

Ferner haben Lord und Novick auch schon gezeigt, dass die *Reliabilität* eines solchen Tests durch die Kuder-Richardson Formel 21 (K – R 21) gegeben ist (Lord & Novick, 1968, S. 523; De Gruijter & Van der Kamp, 1984, S. 60; Klauer, 1987; S. 151 f).

Außerdem steht dann zusätzlich noch die Möglichkeit zur Verfügung, die Zufallsstichproben als eine Art Paralleltests aufzufassen, als zufallsparallele Tests („random parallel tests" bei Lord & Novick) und entsprechend die Zufallsparalleltest-Reliabilität zu ermitteln. In den CBM-Studien sprach man einfach von Paralleltests und Paralleltestreliabilität, was – streng genommen – nicht korrekt ist. Immerhin brachte die so ermittelte Reliabilitätsschätzung bislang im Allgemeinen brauchbare Ergebnisse.

Schließlich kommt als dritte Form auch die sogenannte Split-half-Reliabilität nach der Testhalbierungsmehode in Frage, mit der Strathmann und Klauer (in Vorbereitung) bei der Lernverlaufsdiagnostik in Mathematik besonders gute Erfahrungen gemacht haben. Dabei wird der Test eines jeden Kindes in zwei Hälften aufgeteilt – zum Beispiel in alle geradzahligen Items gegen alle ungeradzahligen Items –, so dass die Anzahl der Lösungen in der einen Hälfte mit der Anzahl der Lösungen in der anderen Hälfte über alle Probanden hinweg korreliert wird. Diese Korrelation ergibt aber erst eine angemessene Schätzung der Reliabilität, wenn sie gemäß Spearman-Brown aufgewertet wird. Auch wenn die Kinder immer neue Itemstichproben erhalten, lässt sich diese Art der Reliabilität dennoch problemlos schätzen.

Auf der Basis von Itemsampling und binomialem Testmodell bieten sich also Möglichkeiten, die wichtigsten Probleme der Lernverlaufsdiagnostik in den Griff zu bekommen, was nun kurz erläutert werden soll.

Konsequenzen

Als offene Probleme der Lernverlaufsdiagnostik waren diese herausgestellt worden: Homogene Testschwierigkeiten, Validität der Tests, das Testmodell und die Änderungssensibilität der Lernverlaufsdiagnostik.

Die Frage nach der Validität der Tests und die nach dem Testmodell lassen sich am einfachsten beantworten. Sind die Aufgabenmengen präzise definiert und die Aufgabenstichproben nach den hier kurz dargestellten Verfahren erzeugt worden, so handelt es sich in jedem Fall um kontentvalide Tests. Streng genommen kann Kontentvalidität nur durch die Erzeugungsprozedur der Tests garantiert werden. Andere Aspekte der Validität wie der Kriteriumsvalidität oder der faktoriellen Validität sind damit nicht erfasst, sondern bedürfen der ergänzenden Prüfung. Allerdings war es aber gerade auch die Kontentvalidität, die vielfach doch problematisch war und jedenfalls nicht systematisch abgesichert werden konnte. Sie wird nun durch die Herstellungsprozedur gewährleistet.

Die Frage nach dem Testmodell ist eindeutig und klar gelöst, wenn man die Aufgabenmengen entsprechend definiert hat und daraus nach geeigneten Vorgaben Zufallsstichproben für jedes Kind und jeden Testter-

min zieht. Beim Itemsampling haben wir es mit dem bewährten binomialen Testmodell zu tun, das Antworten auf alle testtheoretischen Fragen zu bieten hat.

Etwas komplexer stellt sich die Frage nach der Homogenität der Testschwierigkeiten dar. Typischerweise sind die einzelnen Aufgaben in schulisch relevanten Bereichen ungleich schwierig. Selbst wenn man Zufallsstichproben von Aufgaben aus einer definierten Menge zieht, ist keineswegs garantiert, dass diese gleich schwierig sind. Man hat sicher gute Chancen, wenn die Aufgabenstichproben hinreichend groß sind, doch wäre im Einzelnen zu prüfen, wann dies der Fall ist. Das hängt auch von der Größe der Grundgesamtheit ab: Bei 35 Einmaleinsaufgaben hat man es sicher leichter als bei einem Rechtschreibtest auf der Basis eines Grundwortschatzes von etwas über 1300 Wörtern (Strathmann, Klauer & Greisbach, 2010). Hier helfen nur entsprechend große Stichproben oder Stichproben aus Grundmengen, die in homogene Teilmengen nach dem Muster von Tabelle 1 gegliedert und für die proportional-stratifizierte Festlegungen vorgegeben sind, wie man die Teilmengen in den Stichproben zu repräsentieren hat. Falls dies angemessen durchgeführt wird, werden auch kleinere Itemstichproben homogen schwierig sein, wie dies bei Strathmann und Klauer (2010) in der Grundschulmathematik der Fall war (vgl. auch Strathmann & Klauer, in Vorbereitung).

Solange bei alternativen Vorgehensweisen nicht weitere Erfahrungen vorliegen, wird es angebracht sein, die Schwierigkeitshomogenität der Tests empirisch nachzuweisen. Der experimentelle Nachweis der Änderungssensibilität ist – soweit ich sehe – noch eine Aufgabe der Zukunft.

Exkurs zur Änderungssensibilität

Das hier skizzierte Vorgehen ermöglicht, beliebig viele kontentvalide Itemstichproben zu ziehen, die stets die gleiche Kompetenz erfassen. Auf der Basis solcherart erzeugter Itemstichproben kann man daran gehen, die Änderungssensibilität der Tests experimentell nachzuweisen. Wichtig ist dabei, den Zwei-Gruppen-Versuchsplan einzusetzen: Erst wenn es eine Kontrollgruppe ohne spezielle Förderung gibt, können *Retesteffekt* und *Regressionseffekt* kontrolliert werden. Der Retesteffekt besteht in einer Verbesserung der Testleistung aufgrund der vorausgegangenen Testerfahrung. Er kann unabhängig davon auftreten, ob sich der zu messende Personenparameter geändert hat oder nicht. Dasselbe gilt auch für den Regressionseffekt. Das soll kurz erläutert werden.

Tack (1986) hatte schon vor Jahren nachgewiesen, dass klassisch konstruierte Tests praktisch nicht geeignet sind, Änderungen zu erfassen. Zum einen lässt sich damit der Regressionseffekt nicht ausschließen und zum andern kann man das Reliabilitäts-Validitäts-Dilemma im Rahmen der klassischen Testtheorie nicht überwinden. Die Regression zur Mitte ist ein empirisch belegbarer Effekt und besteht darin, dass Probanden, die etwas extremere Werte hatten, dazu neigen, beim Retest Werte zu zeigen, die näher am Mittelwert liegen: Gute erzielen beim Retest tendenziell schwächere Ergebnisse, Schwache beim Retest tendenziell etwas bessere, ohne dass sich die Leistungen verändert haben müssen. Erklären lässt sich das Phänomen durch die Verteilung der Messfehler: Extrem große oder extrem kleine Messfehler treten in der Regel seltener auf, so dass beim Retest der einzelne eher einen weniger extremen Messfehler erhält: Wer einen großen (positiven oder negativen) Messfehler „erwischt" hatte, wird beim nächsten Mal sehr wahrscheinlich einen kleineren Messfehler aufweisen, also etwas näher zum Mittelwert rücken. Wer aber beim ersten Test einen minimalen Messfehler erhalten hatte, wird beim zweiten Test einen etwas größeren Fehler bekommen. Im Ergebnis werden sich manche scheinbar verschlechtern, andere scheinbar

verbessern. Werden allerdings Tests über einen längeren Zeitraum wiederholt gegeben wie bei einer Lernverlaufsdiagnostik, so bestehen gute Chancen des Ausgleichs, so dass der Regressionseffekt zwar die Variabilität der Messwerte erhöht, nicht aber deren längerfristige Tendenz verfälscht.

Anders liegen die Dinge bezüglich des Reliabilitäts-Validitäts-Dilemmas. 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 sich alle Probanden beim nächsten Test um zwei Punkte verbessert, so resultiert eine Korrelation zwischen den beiden Tests von r = 1. Haben die Probanden aber unterschiedlich zugelegt und manche sich sogar verschlechtert, so müssen niedrigere Retestreliabilitäten resultieren. Änderungssensible Verfahren sollten daher nur *mäßige* Retest- oder Paralleltestreliabilitäten aufweisen, aber hohe Werte der Split-half Reliabilität. Außerdem kann man bei Testwiederholung nicht sicher sein, dass der Test nach wie vor dasselbe misst – es sei denn, wir haben es mit zufallsparallelen Tests zu tun, bei denen die Herstellungsprozedur die Kontentvalidität gewährleistet.

Von daher lassen sich folgende Vorhersagen ableiten:

Voraussage 1: Wenn ein Test Änderungen erfasst, so bietet er nur mittelhohe Retest- oder Paralleltestreliabilitäten. Das gilt auch dann, wenn es sich um immer neue Tests im Sinne des Itemsamplings handelt.

Voraussage 2: Wenn ein Testverfahren Änderungen über einen längeren Zeitraum erfasst, so wird die Korrelation zwischen den Tests mit der Länge des jeweiligen Zeitraums immer kleiner. Je länger der erfasste Lernzeitraum, desto niedriger werden die Korrelationen zwischen den Tests ausfallen. Voraussage 2 gilt auch für zufallsparallele Tests.

Voraussage 3: Wenn ein Testverfahren Änderungen erfasst, so bringt das Split-half-Verfahren höhere Reliabilitäten als Reliabilitätsschätzungen mittels Retest oder Paralleltests.

Diese Voraussagen können, basierend auf über 3.500 Kindern und 20 Testerhebungen in Mathematik, wie folgt überprüft werden (vgl. Strathmann & Klauer, in Vorbereitung).

Voraussagen 1 und 2:
Mittelwerte M der Zufallsparalleltest-Reliabilität bei Tests im Abstand von
- 2 Wochen
  M = 0.73, N = 19 Korrelationen
- 15 bis 20 Wochen
  M = 0.44, N = 19 Korrelationen

Voraussage 3: Split-half Reliabilität:
- M = 0.86, N = 1.787 Kinder (zur Schuljahrsmitte)
- M = 0.82, N = 1.683 Kinder (zum Schuljahrsende)

Wie man sieht, sind empirisch genau diese Werte nachzuweisen, wie sie theoretisch vorhersagbar sind. Die zweite der drei Vorhersagen lässt sich noch besser belegen, wenn man wie in Abbildung 1 die zeitlichen Abstände differenzierter berücksichtigt, die zwischen den Tests lagen. Alle zwei Wochen wurde über ein ganzes Schuljahr hinweg immer ein Test erhoben. Die Abbildung zeigt die mittleren Korrelationen zwischen dem ersten Test und den nachfolgenden 19 Tests, die alle dieselbe Kompetenz messen.

Abbildung 1 verdeutlicht den Reliabilitätsabfall der zufallsparallelen Tests im Laufe eines ganzen Schuljahres. Die Abbildung wäre nicht korrekt interpretiert, würde man schließen, dass die Tests im Laufe des Schuljahres markant an Reliabilität verlören. Das ist keineswegs der Fall, wie aus den Split-half-Reliabilitäten hervorgeht. Abbildung 1 lässt aber darauf schließen, dass die Kinder im Laufe des Schuljahres *unterschiedlich* viel gelernt haben, dass sie sich in den *Zuwächsen* deut-

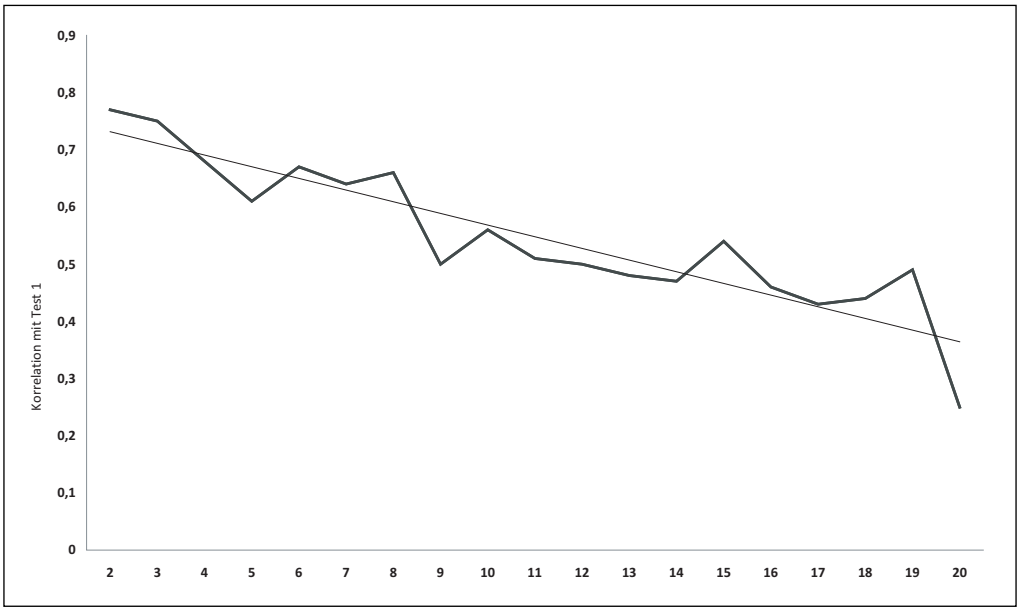

Abb. 1: Korrelation des ersten Tests mit den 19 folgenden Tests im Verlauf des Schuljahres

lich unterscheiden. Das ließe sich leicht auch an den individuellen Lernverläufen demonstrieren, worauf unten nur kurz eingegangen werden soll.

Diese Befunde sprechen eindeutig dafür, dass es sich um änderungssensible Testverfahren handelt. Die Daten sollten dazu ermutigen, auch noch den experimentellen Nachweis zu führen.

Lernverlaufsdiagnostik

Zwei Varianten von Lernverlaufsdiagnostik

Offenbar kann man im pädagogisch-psychologischen Kontext mindestens zwei verschiedene Formen von Verlaufsdiagnostik unterscheiden. Angenommen, es soll eine bestimmte Fertigkeit, die im Prinzip zwar schon beherrscht wird, nun durch Übung verbessert werden, so dass die Aufgaben *schneller* bewältigt oder/und *weniger Fehler* gemacht werden. Man denke etwa an ein Training der Aufmerksamkeit, bei dem die Leistungen im Aufmerksamkeitstest am Schluss des Trainings schneller bearbeitet werden und weniger Fehler anfallen sollen. Oder man denke an die Übung der Lesefertigkeit, so dass die Kinder flüssiger lesen und weniger Lesefehler produzieren. In solchen Fällen wird man immer Tests bringen, die prinzipiell die gleiche Fertigkeit erfassen, um so fortlaufend die Verbesserungen in Form abnehmender Fehler oder zunehmender Lösungsgeschwindigkeit dokumentieren zu können. Das ist beispielsweise beim Lesetest von Walter (2010a) der Fall.

Geht es aber darum, das Wissen und Können nicht nur zu verbessern, sondern substantiell zu *erweitern*, so kann man den Fortschritt nur erfassen, wenn von Anfang an geprüft wird, was am Ende erreicht sein soll. In der Lernverlaufsdiagnostik wird man also mindestens zwei Varianten unterscheiden müssen,

(a) die Verbesserung einer Kompetenz bezüglich Geschwindigkeit und/oder Genauigkeit und

(b) die Erweiterung einer Kompetenz um immer neue Komponenten.

In der Mathematik und in den meisten Fächern haben wir es typischerweise mit dieser zweiten Variante zu tun. Und um den Lernverlauf zu dokumentieren, ist in dem Fall jeweils zu erfassen, wie gut das gekonnt wird, was *am Ende* gekonnt sein soll. Das führt zu speziellen Konsequenzen, die im Folgenden deutlicher werden.

Neu: Lernverlaufsdiagnostik Mathematik Grundschule

Nach den Grundsätzen, die oben für lehrzielorientierte Tests auf der Basis des Binomialmodells vorgestellt worden sind, wird demnächst ein Programm verfügbar sein, das die Lernverlaufsdiagnostik in Mathematik für die Grundschulklassen 2, 3 und 4 auf eine Weise ermöglicht, an die bislang nicht zu denken war (Strathmann & Klauer, in Vorbereitung). Das Programm kann so auf einen PC geladen werden, dass Lehrer beliebig oft und beliebig viele Testblätter ausdrucken können, wobei keine zwei Blätter identisch sein werden, aber alle dasselbe Lehrziel abdecken. Das soll nun im Einzelnen etwas näher erläutert werden.

Die „Lernverlaufsdiagnostik Mathematik" wird wie Walters „Lernfortschrittsdiagnostik Lesen" in der Reihe Deutsche Schultests erscheinen und aus einem Manual sowie einer CD bestehen. Das Manual bietet zunächst eine Einführung in das Konzept der Verlaufsdiagnostik und beschreibt dann differenziert die Definition der Aufgabenmengen. Für jedes Schuljahr werden die Aufgabenmengen der vier Operationen Addition, Subtraktion, Multiplikation und Division in der Art definiert, wie dies in Tabelle 1 für die Subtraktion im 2. Schuljahr geschehen ist. Insgesamt werden pro Testtermin 24 Aufgaben ausgedruckt, je sechs für die vier Operationen.

In der CBM-Tradition hat sich bewährt, relativ kurze Tests zu geben, die also wenig Zeit in Anspruch nehmen, so dass die Tests auch sehr oft eingesetzt werden können, ohne die Lernphasen stärker zu beeinträchtigen. Daher also die Beschränkung auf 24 Items, zumal es bei dieser Anzahl möglich ist, die vier Operationen gleichmäßig mit je sechs Aufgaben zu berücksichtigen.

Die Definitionen der einzelnen Aufgabenmengen enthalten in der Regel mehr als sechs Teilmengen, die also nicht alle berücksichtigt werden können. Deshalb sind pro Operation und Schuljahr auch spezielle Anweisungen im Programm eingebaut, Generierungsregeln, so dass bei jeder Erzeugung von Aufgabenblättern die Teilmengen stets in den gleichen Proportionen repräsentiert werden. Es handelt sich also um ein Verfahren des proportional-zufälligen Itemsamplings. Nur auf solche Weise lässt sich gewährleisten, dass zwar mittels Zufallsgenerator immer neue Aufgaben gebildet werden, die verschiedenen Aufgabenblätter aber stets die Grundmenge in der gleichen Weise repräsentieren. Durch eine solche Prozedur lässt sich die Erzeugung gleich schwerer Aufgabenblätter realisieren. Konkret konnte dieser Aspekt empirisch bestätigt werden.

Das Manual bietet des Weiteren Anweisungen zur Durchführung der Tests und zur Auswertung. Bei der Auswertung stößt man auf den *ersten* Nachteil des neuen Verfahrens: Jedes Kind bekommt ja ein eigenes Testblatt, damit das Binomialmodell anwendbar ist, so dass jedes Aufgabenblatt eigens ausgewertet werden muss. Manche vermuten zwar, dass das eigene Blatt pro Kind nur das Abschreiben ausschließen soll, aber das ist ein unbeabsichtigter Nebeneffekt. Tatsächlich geht es um die Anwendbarkeit des bewährten Testmodells. Man kann die bearbeiteten Aufgabenblätter zwar selbst nachrechnen, aber das Programm bietet zwei bedeutsame Hilfen an: Man lässt sich pro Kind ein Blatt mit den Lösungen ausdrucken oder gibt die Lösungen des Kindes in den PC selbst

ein, wobei der Computer die Auswertung übernimmt.

Es gibt aber noch einen *zweiten* Nachteil des Verfahrens: Lernverlaufsdiagnostik erfordert hier, dass schon vom Schuljahresbeginn an getestet werden muss, was erst am Schuljahrsende gekonnt sein soll. Es werden also von Anfang an 24 Aufgaben über *alle* Teilziele hinweg den Kindern vorgelegt, obwohl sie bei weitem noch nicht alle lösen können. Das lässt sich nicht vermeiden, wenn der Lernverlauf wirklich abgebildet und etwa in Verlaufskurven dargestellt werden soll. Wie man sieht, hat das Mädchen von Abbildung 2 etwa bis zum siebten Test nur um die 4 – 6 Aufgaben richtig gelöst, begann dann aber, möglicherweise in Abhängigkeit vom Fortschritt des Unterrichts, sich recht kontinuierlich zu verbessern.

Wird jedes Mal komplett das abgetestet, was das Lehrziel für das Schuljahr ist, so werden am Schuljahresbeginn nur wenige der geforderten 24 Aufgaben gelöst werden können, aber der Anteil lösbarer Aufgaben wird ansteigen. Um die Kinder nicht zu entmutigen, ist es sinnvoll, ihnen zu sagen, dass sie manche Aufgaben noch gar nicht rechnen können und dass sie diese einfach auslassen dürfen. So sind insbesondere in den ersten Wochen des Schuljahres nur wenige Aufgaben pro Kind auf Korrektheit zu prüfen. Das macht auch Abbildung 2 deutlich.

Die bemerkenswerte Lernverlaufskurve des Mädchens von Abbildung 2 wäre nicht möglich gewesen, wenn jedes Mal eben nur jene Aufgabentypen gestellt worden wären, die die Kinder jetzt lösen können sollten, wie dies in der Regel bei Klassenarbeiten vorgesehen ist. Würde man bei solchen Arbeiten jedes Mal den Prozentsatz gelöster Aufgaben berechnen, so würde der bei den meisten Kindern um einen Mittelwert variieren und den Fortschritt nicht erkennen lassen. Verläufe wie in Abbildung 2 könnten dann normalerweise nicht gesehen werden.

Die Lernverlaufskurven zeigen stattdessen, wo das einzelne Kind oder die einzelne Klasse mit Blick auf das Lehrziel des Schuljahres derzeit steht, welcher Anteil gerade geschafft wird und wie viel noch fehlt. Abbildung 3 bietet im Kontrast zu Abbildung 2 die Leistungsentwicklung einer vierten Klasse Grundschule über ein ganzes Schuljahr hinweg. Diese Grafik ist deshalb besonders be-

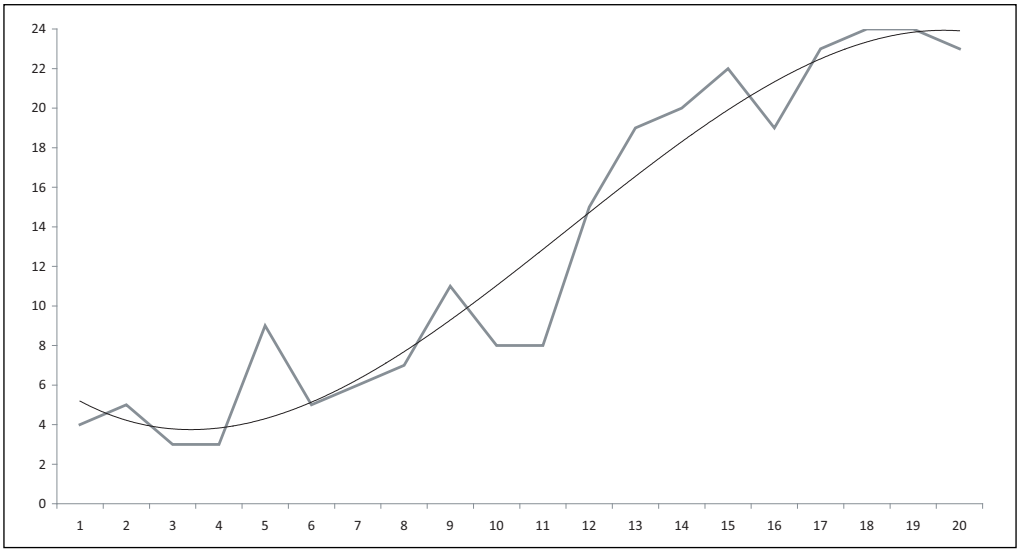

Abb. 2: Verlaufskurve eines Mädchens der 2. Klasse

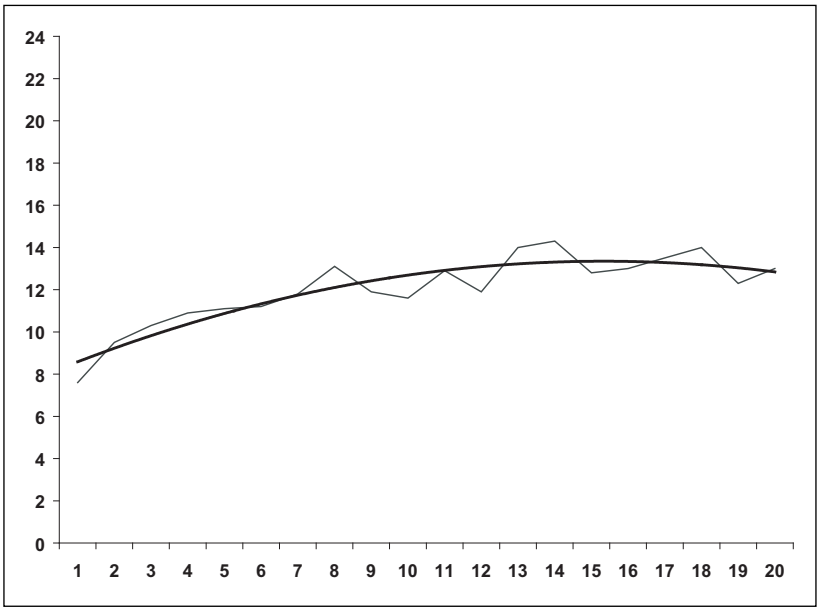

Abb. 3: Lernverlaufskurve einer vierten Klasse Grundschule

merkenswert, weil die Kinder am Schuljahresbeginn schon rund ein Drittel der Aufgaben beherrschten, die sie am Ende können sollten, dann aber kaum mehr etwas dazulernten. Der Lernzuwachs schwächte sich sogar im Laufe der Zeit etwas ab, und man fragt sich, ob der Lehrkraft diese Entwicklung überhaupt zur Kenntnis gelangt ist. Immerhin startete die Klasse mit einem guten Leistungsniveau, hatte also im vorigen Schuljahr erfolgreich gelernt, was in diesem Schuljahr nun keineswegs der Fall war. Woran das wohl gelegen haben mag?

Abschließend sei noch kurz auf die *Normen* eingegangen, die das Manual bietet.

Grundsätzlich ermöglicht das Programm die Anwendung dreier Arten von Normen, lehrzielorientierte Normen, individuelle und soziale Normen. Im Lernverlaufstest Mathematik geht es bei jeder Anwendung zunächst darum zu prüfen, welcher Anteil vom Lehrziel inzwischen beherrscht wird. Angemessen sind demnach *lehrzielorientierte Normen*, die das Manual auch auf der Basis des Binomialmodells anbietet. Diese Normen bestehen aus dem Prozentsatz richtig gelöster Aufgaben und offerieren zusätzlich das zugehörige Konfidenzintervall. Testtheoretisch handelt es sich dabei also um die aktuelle Schätzung des Personenparameters des jeweiligen Kindes.

Tatsächlich kann man auf dieser Grundlage auch lehrzielorientierte Noten geben (vgl. Klauer, 1987, S. 294-295), doch wird im Manual darauf verzichtet, weil die amtliche Notengebung nicht streng lehrzielorientiert vorgeht.

Interessierte Pädagogen können aber bei dem neuen Verfahren auch die *individuelle Norm* anwenden. Man wird dann Aussagen darüber geben, ob sich das Kind verbessert oder ob es sein Niveau wenigstens gehalten hat oder ob es sich gar verschlechterte. Aussagen dieser Art sind bei herkömmlichen Klassenarbeiten nicht möglich. Dort bietet sich nur der *soziale Vergleich* mit den anderen Kindern in der Klasse an, bei dem in aller Regel immer dieselben Kinder oben rangieren und andere stets unten.

Die Anwendung der individuellen Norm kann das Selbstkonzept der Kinder stützen, aber auch vor überzogenem Selbstkonzept

bewahren. Angenommen, ein Kind löst beim ersten Test nur eine Aufgabe richtig, bekommt alle zwei Wochen einen weiteren Test und verbessert sich dabei stetig um etwa eine weitere Aufgabe, so erhält sein Selbstkonzept jedes Mal Auftrieb; betrachtet man den Lernverlauf in der grafischen Darstellung, so steigt die Kurve in schöner Regelmäßigkeit an: Dennoch bleibt das Kind am Ende des Schuljahres noch deutlich unterhalb des angestrebten Lehrziels seiner Klasse, leistet also nicht, was es leisten sollte.

In der Reihe „Deutsche Schultests" ist es darüber hinaus üblich, bundesweite Vergleichsnormen anzubieten. Diese spezielle Art *sozialer* Normen bietet das Manual ebenfalls, obwohl sie nicht zum Konzept der Lernverlaufsdiagnostik gehören, denn solche Normen beziehen sich nicht auf den Lernverlauf, sondern auf bestimmte Testzeitpunkte. Die „Lernverlaufsdiagnostik Mathematik" offeriert zwei Möglichkeiten zu bundesweiten Vergleichen, nämlich jeweils zur Mitte und zum Ende des Schuljahres. Bei der Eichung wurden fast 3500 Kinder herangezogen, wobei sich die Anteile der einzelnen Bundesländer daran orientierten, welchen Anteil sie an der Gesamtheit aller Grundschüler aufzuweisen haben.

Schließlich bietet das Manual ausführliche Informationen über die Untersuchungen zu den Gütekriterien des Tests, beginnend bei der Prüfung der Homogenität der Testschwierigkeiten und weiterführend zur Reliabilität und Validität, die ebenfalls mehrfach und mit verschiedenen Ansätzen überprüft worden sind. Hierzu liegen in allen Bereichen recht befriedigende Ergebnisse vor.

Die CD bietet die Möglichkeit, beliebig viele und beliebig oft immer neue Aufgabenblätter für jedes Kind auszudrucken – wenn man will gleich auch separat die Lösungen. Ab dem fünften Test kann man darüber hinaus für jedes Kind eine Graphik des Lernverlaufs ausdrucken und natürlich auch den Lernverlauf der ganzen Klasse. Das setzt allerdings die Eingaben der Daten voraus, was automatisch geschieht, wenn man die Blätter, die die Kinder bearbeitet haben, am PC auswertet. Außerdem eröffnet das Programm die Möglichkeit, die Daten der Klasse zu verwalten und nach verschiedenen Aspekten hin zu analysieren.

Literatur

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

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

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

Gross, A. L. & Shulman, V. (1980). The applicability of the beta binomial model for criterion-referenced tests. Journal of Educational Measurement, 17, 195-202.

Grosche, M. & Hintz, A.-M. (2010). Überprüfung von Verfahren zur Evaluation von Alphabetisierungskursen durch eine Einzelfallstudie. Heilpädagogische Forschung, 36, 177-185.

Kempf, W. F. (1977). A dynamic test model and its use in the micro-evaluation of instructional material. In H. Spada & W. F. Kempf (Hrsg.). Structural models of thinking and learning. Bern: Huber.

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

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

Krauth, J. (1995). Testkonstruktion und Testtheorie. Weinheim: BeltzPVU.

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

Petermann, F. (1978). Veränderungsmessung. Stuttgart: Kohlhammer.

Petermann, F. (2010). Veränderungsmessung. In D. H. Rost (Hrsg.), Handwörterbuch Pädagogische Psychologie (S. 896-902). Weinheim: Beltz.

Petermann, F. & Hehl, F.-J. (1979). Einzelfallanalyse. München: Urban & Schwarzenberg.

Roskam, E. E. (1996). Latent-Trait-Modelle. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg). Handbuch Quantitative Methoden (S. 430-458). Weinheim: BeltzPVU.

Rost, J. & Spada, H. (1983). Die Quantifizierung von Lerneffekten anhand von Testdaten. Zeitschrift für Differentielle und Diagnostische Psychologie, 4, 29-49.

Scheiblechner, H. (1996). Item-Response-Theorie: Prozessmodelle. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg), Handbuch Quantitative Methoden (S. 459-466). Weinheim: BeltzPVU.

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

Strathmann, A. M. & Klauer, K. J. (2010). Lernverlaufsdiagnostik: Ein Ansatz zur längerfristigen Lernfortschrittsmessung. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 42, 111-122.

Strathmann, A. M. & Klauer, K. J. (in Vorbereitung). Lernverlaufsdiagnostik Mathematik. Göttingen: Hogrefe.

Strathmann, A. M., Klauer, K. J. & Greisbach, M. (2010). Lernverlaufsdiagnostik. Dargestellt am Beispiel der Entwicklung der Rechtschreibkompetenz in der Grundschule. Empirische Sonderpädagogik, 2, 64-77.

Tack, W. H. (1986). Reliabilitäts- und Effektfunktionen – ein Ansatz zur Zuverlässigkeit von Messwertänderungen. Diagnostica, 32, 48-63.

Walter, J. (2008). Curriculumbasiertes Messen (CBM) als lernprozessbegleitende Diagnostik: Erste deutschsprachige Ergebnisse zur Validität, Reliabilität und Veränderungssensibilität eines robusten Indikators zur Lernfortschrittsmessung beim Lesen. 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.

Walter, J. (2010a). Lernfortschrittsdiagnostik Lesen. Ein curriculumbasiertes Verfahren. Göttingen: Hogrefe.

Walter, J.(2010b). Lernfortschrittsdiagnostik am Beispiel der Lesekompetenz (LDL): Messtechnische Grundlagen sowie Befunde über zu erwartende Zuwachsraten während der Grundschulzeit. Heilpädagogische Forschung, 36, 162-176.

Anschrift des Autors

Prof. em. Dr. Karl Josef Klauer
RWTH Aachen
Privat: Robert-Stolz-Weg 15
42781 Haan
klauerk@uni-duesseldorf.de