Thomas Lerche

Leistungsbeurteilung an Schulen

Lehrstuhl für Schulpädagogik,
Ludwig-Maximilians-Universität München

April 2022

Inhaltsverzeichnis

Einleitung – Leistungsbeurteilung als professionelle Handlung

Die Beurteilung und Benotung von Schülerinnen und Schülern ist seit jeher ein wesentlicher Bestandteil der europäischen Schulsysteme. Zugleich handelt es sich dabei um einen der komplexesten Prozesse des Lehrerhandelns, denn er findet in einer von Widersprüchen geprägten Umgebung statt, in der sich rechtliche Vorgaben, Forderungen kritischer Eltern und Ansprüche der Gesellschaft zu einem schwer durchschaubaren Wust an Anforderungen vermischen können. Nicht zuletzt deswegen gilt dieser Bereich als besonders angstbelastet, als etwas, bei dem gerade junge Lehrkräfte eine große Unsicherheit verspüren (Jürgens 2005; Jerusalem/Mittag 1999). Dazu tragen auch zirkulierende Mythen von Schulleitern bei, die bestimmte Notenschnitte verlangen, von klagewütigen Eltern und ungerechten Bewertungen. Dieses Buch soll Ihnen helfen, die Herausforderungen auf diesem Gebiet zu meistern, indem es Sie mit verschiedenen Situationen bekannt macht, Ihnen theoretisches, rechtliches und angewandtes Wissen vermittelt und zu Ihrer Professionalisierung beiträgt – damit Sie das volle Potential schulischer Leistungsbewertung ausschöpfen und Ihre Beurteilungen gegenüber Kollegen, Eltern oder Vorgesetzen jederzeit transparent und fachgerecht vertreten können.

Es ist nicht einfach, eine Lösung für die umfangreichen und schwierigen Herausforderungen der schulischen Leistungsbeurteilung zu finden, zumal alle Anforderungen (inklusive der Gesetze und Verordnungen) normativ formuliert sind und somit nur schwer für den Lehreralltag funktionalisiert werden können. Schlömerkemper fasst das zugrundeliegende Problem sehr gut zusammen, wenn er betont,

„wie schwierig es ist, die Bedeutung von Leistung und Leistungsmessung angemessen einzuschätzen und pädagogisch sinnvoll damit umzugehen. Dies ist eben nicht generell und allgemeingültig zu entscheiden und zu handhaben, sondern jeweils nur in aktuellen Situationen der Interaktion zwischen den jeweils Betroffenen zu bestimmen" (Schlömerkemper 2002, S. 319).

Es ist also nicht möglich, Handlungskompetenzen zu definieren, die in jeder, auch in der komplexesten, Situation anwendbar sind. Frei nach Gieseke (2007) formuliert: Es gibt keine richtige Anwendung der Leistungsdiagnostik, sondern nur eine angemessene.

Der geeignete Ansatzpunkt zur Bewältigung der unterschiedlichen und unsicheren Herausforderungen des Lehrberufs – und damit auch der schwierigen Aufgabe der Leistungsbeurteilung – ist bereits über 200 Jahre alt und stammt von Johann Friedrich Herbart (1806/1986). Forschungsanstrengungen der heutigen Zeit, die zum selben Thema gemacht werden, wie beispielsweise die Professionalisierungstheorie (Bauer/Kopka/Brindt 1996) oder die Expertiseforschung (Ericsson/Krampe/Tesch-Römer 1993; Bromme 1992) kommen allesamt zu derselben Schlussfolgerung: Der Lehrberuf ist im Kern eine Aufgabe, die nach professionellem Wissen und Handeln verlangt. Er fordert Expertise, also auf einem bestimmten Gebiet dauerhaft (nicht zufällig oder singulär) herausragende Leistungen zu erbringen (Gruber 1994). Der Weg zur professionellen Expertise ist langwierig, kann nicht deterministisch vorgeplant werden und verlangt Ernsthaftigkeit, theoriebasierte Wohldurchdachtheit, Berücksichtigung der berufspraktischen Erfahrungen und Fleiß. Herbart nennt diese Expertise den Pädagogischen Takt und bezeichnet damit, in einfachen Worten, das pädagogische Gespür für die Situation.

Für Herbart sind hierbei zwei Aspekte wichtig:

- Zum einen ist es für eine Lehrkraft nachteilig, wenn ihr Handeln in Routinen ohne Reflexion mündet. Den alleinigen Bezug auf die praktischen Erfahrungen bezeichnet Herbart als „Schlendrian eines ewig gleichförmig handelnden Schulmeisters, der nichts erfährt, weil er nur sich erfährt" (Herbart 1986, S. 243). Erfahrungen der Praxis, die theorieblind bleiben, können nicht für ein sachangemessenes Handeln verwertet werden.
- Zum anderen darf die Lehrkraft nicht davon ausgehen, dass eine allgemeingültige Theorie das Erziehungshandeln, ähnlich wie etwa die Bedienung technischer Geräte, vollständig funktional normieren könnte.

Das professionsspezifische Gespür bildet sich weder unmittelbar durch den reinen Erwerb theoretischen Wissens noch allein durch pädagogische Praxis. Es entsteht vielmehr als dialektische Bewegung zwischen Theorie und Praxis, d. h. durch wohldurchdachtes Handeln im Berufsalltag. Die im Umgang mit konkreten Herausforderungen gemachten Erfahrungen müssen vor einem theoretischen Hintergrund reflektiert, gegen andere Handlungsalternativen abgewogen und im Hinblick auf zukünftige Situationen bewertet werden. So erwirbt man durch langjährige und ausdauernde theoriebasierte Auseinandersetzung mit eigenen Erfahrungen im schulischen Problemraum ein gutes Gespür für die passende Reaktion auf unvorhergesehene Zwischenfälle, den Pädagogischen Takt. Herbart selbst machte keine Zeitangabe, wie lange dieser Lernprozess dauert, die moderne Expertiseforschung (Ericsson/ Krampe/Tesch-Römer 1993) nennt einen Zeitraum von ungefähr 10 Jahren bzw. 10 000 Stunden: „Ten thousand hours is the equivalent to roughly three hours per day, or twenty hours per week, of practice over ten years“ (Levitin 2006, S. 197).

Das Ziel einer jeden Ausbildung ist es demnach, dass ihre Absolventen die Problemstellungen dieser Wissensdomäne eigenständig und erfolgreich lösen können. Neben dem genannten hohen Maß an Faktenwissen benötigen sie dafür (Gruber/Stöger 2011):

- reichhaltige Erfahrungen mit Problemstellungen der Wissensbasis,
- hohe Fähigkeiten zur Diagnose und der Bearbeitung von Problemen,
- hohe metakognitive Fähigkeiten,
- hohe Sicherheit in der Anwendung von Lösungsstrategien und Heuristiken sowie
- große Flexibilität gegenüber neuen Problemsituationen.

In den Schulalltag übertragen bedeutet dies:

- Sie müssen wissen, was Sie tun.
- Sie müssen Ihr Handeln pädagogisch begründen.
- Sie müssen pädagogische Diagnostik so anwenden, dass sie der Förderung des Lernprozesses dient.
- Sie müssen über die notwendigen pädagogischen Handlungen mit Fachleuten (z. B. der Schulleitung) und mit Laien (z. B. Eltern) lösungsorientiert, nachvollziehbar und verständlich diskutieren können.

Ein wichtiger Ausgangspunkt der Professionalisierung ist zunächst einmal grundlegendes theoretisches Wissen, um eine angemessene Basis für die Reflexion der Praxiserfahrungen zur Verfügung zu haben (dies ist übrigens der wesentliche Grund für theorielastige Ausbildungsabschnitte: Sie benötigen hinreichend viel theoretisches Wissen, um Ihre Praxiserfahrungen produktiv reflektieren zu können). Diese Wissensbasis wird im Laufe des Expertiseerwerbs kontinuierlich angepasst, beispielsweise wenn für eine bestehende Anforderung das regelgeleitete, theoretisch begründete Rüstzeug nicht ausreicht, sondern neue, angemessene Lösungen gefunden werden müssen, mit denen die Widersprüche überwunden werden können (vgl. Modell der Kompetenzentwicklungsphasen nach Dreyfus/Dreyfus 1986; Berliner 2001; Keller-Schneider 2010, S. 60). Einen möglichen Baustein für solche Anpassungen liefern die theoretischen Abhandlungen dieses Buches.

Nun zeigen Forschungen zum Expertiseerwerb allerdings, dass reines Auswendiglernen von Fakten nicht ausreicht. Neben den bereits angesprochenen Reflexionen der Praxiserfahrungen vor einem pädagogischen Hintergrund helfen dabei – gerade in Ausbildungsabschnitten ohne reichhaltige Praxis – authentische Fälle aus der Unterrichtswirklichkeit zum Thema Leistungsbeurteilung. Aus diesem Grund beginnt jedes der Kapitel mit drei zum Thema passenden Episoden, die während des Kapitels, ergänzt um die theoretischen Wissensbereiche, erneut aufgenommen werden.
Dieses Buch kann kein Gesamtwerk über alle Facetten der Leistungsbeurteilung in der Schule sein. Dafür sind die möglichen Fälle zu vielfältig und unübersichtlich. Die Aufgabe dieses Buches besteht vielmehr darin, die Gemeinsamkeiten und Schnittpunkte der Theorien der empirischen Sozialforschung, der

Testtheorie, der Theorie des Lehrens und Lernens, der pädagogischen Diagnostik und der Leistungsbeurteilung in der Schule so mit der Schulwirklichkeit zu verknüpfen, dass exemplarische Handlungsrichtlinien – gerade für die ersten zehn Jahre des Berufsweges – hergeleitet, übernommen, adaptiert und damit erworben werden können.

1. Warum ist schulische Leistungsbewertung wichtig?

Eine kurzgefasste Antwort auf die Frage, warum schulische Leistungsbeurteilung wichtig ist, könnte folgendermaßen lauten: Weil eine professionelle Unterstützung von Lernprozessen ohne Kenntnis der einzelnen Leistungen, die von den Schülern erbracht werden, schlichtweg nicht möglich ist. Diese Aussage verlangt nach einer Begründung und einer Ausarbeitung der daraus resultierenden Implikationen für die Umsetzung in der Schule.

Die Schwierigkeit, den Zusammenhang zwischen Lernprozessen und Leistung zu verstehen, liegt vor allem in den unterschiedlichen Zugängen aller Beteiligten zu diesem Thema. Für gewöhnlich argumentieren alle Personen, die mit dieser zentralen schulischen Aufgabe in Berührung kommen, aus ihrer eigenen, biografisch und beruflich geprägten Befindlichkeit. Diese ist, neben eigenen Erfahrungen, oft auch von zahlreichen Mythen und Konzepten geprägt, die es schwer machen, eine standardisierte und von allen Betroffenen akzeptierte Regelung für diesen schwierigen und unsicheren Bereich zu finden. Folgende Episoden geben hierfür Beispiele:

Episode 1: Das Elterngespräch

Nach der Korrektur einer Probearbeit stehen in der nächsten Sprechstunde mehrere Eltern vor Ihrem Schreibtisch und beschweren sich über die Qualität Ihrer Korrektur. Auf Nachfrage, worin denn genau die Mängel gesehen werden, antworten die Eltern unterschiedlich. Sie notieren sich die Äußerungen:

- *Warum geben Sie so strenge Noten? Mein Sohn soll später mal auf eine gute Schule gehen (vgl. Kap. 2).*
- *Warum legen Sie den Punkteschlüssel erst nachträglich fest? Das ist nicht erlaubt (vgl. Kap. 4)!*
- *Mein Sohn hat sich doch so verbessert. Warum bekommt er immer noch die schlechteste Note (vgl. Kap. 5)?*
- *Wie kommen Sie darauf, bei einer Gruppenarbeit jedem die gleiche Note zu geben. Mein Patrick hat alles alleine gemacht, die anderen sollen schlechter bewertet werden (vgl. Kap. 8)!*

Episode 2: Die Probe wird einkassiert

Bereits zum zweiten Mal in diesem Schuljahr steht ein Gespräch mit dem Schulleiter über eine Ihrer Probearbeiten an. Dabei eröffnet ihnen der Rektor, dass er erneut eine Klassenarbeit von Ihnen aufheben und eine neue Leistungsüberprüfung anordnen wird, da die von Ihnen vorgelegte Probe die Leistung der Schüler/-innen nicht angemessen testet. Auf Ihre Nachfrage, welche Aspekte denn zu diesem Urteil führen, nennt der Schulleiter:

- *Die Verteilung der Noten ist unnormal! Es gibt vor allem Vierer und Zweier, kaum Dreier (vgl. Kap. 6).*
- *Die Klausur war zu schwer!*
- *Die Fragen waren nicht klar!*
- *Sie haben gar nicht durchgenommen, was Sie in der Klausur gefragt haben!*
- *Wie kommen Sie denn eigentlich zu Ihrem Urteil? Ich habe den Eindruck: Was Ihnen gefällt, bekommt gute Noten, was Ihnen nicht gefällt, bekommt schlechte Noten (die Antworten auf all diese Fragen hängen mit Ihrer Professionalität im Umgang mit Leistungsbeurteilung zusammen, vgl. Kap. 6 und 7)!*

Episode 3: Diagnostik als Förderinstrument

Als engagierte Lehrerin ist es Ihnen ein Anliegen, alle Schülerinnen und Schüler bestmöglich zu fördern. Daher geben Sie sich besondere Mühe und schreiben unter jeden Aufsatz ein ausführliches Feedback. Gerade bei Maria erkennen Sie besonderen Förderbedarf und formulieren daher:

Liebe Maria. Ich sehe, dass Du Dir besondere Mühe gemacht hast, den Lückentext auszufüllen. Manchmal ist es Dir gut gelungen und Du hast die richtige Lösung gefunden. Manchmal auch nicht. Hier solltest Du Dir noch mehr Mühe geben, Dich richtig vorzubereiten. Es ist kein Problem, wenn Dir einige Formulierungen schwerfallen. Oft fällt einem die richtige Lösung unter dem Druck der Klassenarbeit nicht ein. Versuche einfach, Dich besser zu konzentrieren und an die Sachen zu denken, die Du gelernt hast, dann wird es schon.

Frau Müller, eine Ihrer Kolleginnen, sagt jedoch, dieses Feedback sei eine Katastrophe. Hat sie Recht (vgl. Kap. 3)?

Falls Sie noch keine eigenen Erfahrungen mit schulischer Leistungsbewertung gemacht haben, illustrieren die geschilderten Episoden die zentrale Herausforderung dieses Bereiches sehr gut: Ohne professionelle Kompetenz können Sie mit den geschilderten Vorhaltungen nicht produktiv umgehen. Wenn Sie nicht über die theoretische Einbettung und Umsetzung im Schulkontext Bescheid wissen, angemessene Handlungen durchführen und eventuelle Fehler oder Ungenauigkeiten beheben können, handeln Sie nicht mit der gebotenen Qualität. Dies führt zu negativen Konsequenzen für alle Beteiligten. Die in den obigen Episoden genannten Vorwürfe sind massiv und erklären gut, warum neben Unterrichtsstörungen und Elternarbeit vor allem die Leistungsbeurteilung der Aspekt ist, der bei angehenden Lehrerinnen und Lehrern die größten Befürchtungen auslöst: Der Handlungsbereich der schulischen Leistungsbeurteilung ist schwierig, umfangreich und nicht vollständig planbar.

Die damit verbundenen Anforderungen haben größtenteils antinomischen, d. h. widersprüchlichen Charakter, sodass die Tätigkeit häufig mit negativen Emotionen verknüpft und von Lehrpersonen subjektiv als belastend empfunden wird (Helsper 1996):

- Die Gesellschaft stellt an Lehrerinnen und Lehrer den Anspruch, Schülerinnen und Schüler bestmöglich individuell zu fördern und dafür auch die Instrumente der pädagogischen Diagnostik fachgerecht anzuwenden.
- Dem gegenüber steht die Selektions- und Allokationsfunktion der Schule, da die Gesellschaft – sehr pauschalisiert ausgedrückt – von ihr fordert, die guten von den schlechten Schüler/-innen zu trennen, um damit beispielsweise berufliche Laufbahnen zu eröffnen oder zu verwehren.
- Es werden von staatlicher Seite angemessene, transparente und kriteriumsorientierte Maßstäbe verlangt, die der Lehrer unter Berücksichtigung der Gleichbehandlung aller Schülerinnen und Schüler, der Vorgaben der Lehrpläne und seiner pädagogischen Verantwortung festsetzen soll (vgl. hierzu z. B. BayEUG, §52).
- Schüler/-innen auf der anderen Seite – und durchaus auch viele Eltern – wünschen eine möglichst milde Beurteilung. Selbst für die Lehrkraft ist es scheinbar weniger belastend, möglichst keine schlechten Noten zu geben.
- Darüber hinaus sind Einzelfälle denkbar, die zwar – anders als beispielsweise die Lese-Rechtschreibschwäche – gesetzlich nicht geregelt sind, eine Gleichbehandlung bei der Bewertung aber dennoch schwierig machen. Hier sind etwa unterschiedliche körperliche Voraussetzungen im Sportunterricht zu nennen.
- Hinzu kommen die praktischen Herausforderungen an die Leistungsbeurteilung: Testgütekriterien sollen eingehalten werden und die Schulleitung wünscht eine Vergleichbarkeit der Anforderungen über die einzelnen Klassen hinweg. Messtheoretische Grundlagen sind jedoch im Praxiskontext Schule nicht in voller Qualität umzusetzen.

Ein Lösungsansatz für unsichere Problemstellungen ist stets in den Gemeinsamkeiten der unterschiedlichen Wünsche und Erwartungen zu suchen (s. auch Kapitel 2). Das verlangt jedoch zuerst nach einer theoretischen Aufarbeitung der Aspekte *Lernen* und *Leistungsbeurteilung*, damit der Zusammenhang zwischen ihnen und damit die Funktionen der Leistungsbeurteilung im schulischen Kontext richtig erfasst werden können.

1.1. Ein theoretisches Modell der schulischen Leistungsbeurteilung

Zu Beginn der Auseinandersetzung mit diesem Themenbereich muss zunächst der Begriff der schulischen Leistungsbeurteilung definitorisch geklärt, strukturiert und in ein Modell schulischen Lehrens und Lernens eingebunden werden. Die Basis dafür bildet das Modell der pädagogischen Diagnostik von Ingenkamp und Lissmann (2008, S. 16), das – im Rahmen dieses Buchs leicht verändert – im Folgenden anhand seiner wichtigsten Teilaspekte vorgestellt wird.

Ziel der schulischen Leistungsbeurteilung

Jede pädagogische Handlung ist auf ein Ziel zu beziehen – das gilt auch für die Leistungsbeurteilung. Im schulischen Kontext lautet dieses Ziel zumeist *Erwerb von Wissen und Kompetenzen bei den Schülerinnen und Schülern.* Nimmt man den Lehrenden als erste Bezugsgröße hinzu, so definieren Ingenkamp und Lissmann (ebd.) folgende Ausgangslage der pädagogischen Diagnostik:

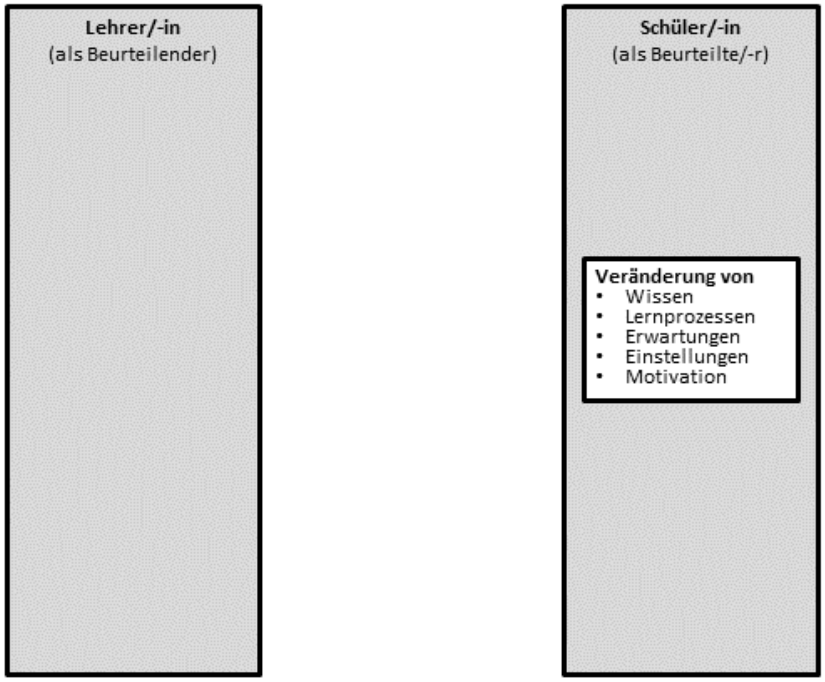

Abb. 4: Modell der pädagogischen Diagnostik (Schritt 1: Das Ziel diagnostischen Handelns)

In der pädagogischen Diagnostik gibt es zunächst zwei Hauptebenen: Den Beurteilenden (hier: Lehrer/-in) und den zu Beurteilenden (hier: Schülerinnen und Schüler). Auch das Ziel ist bereits definiert und wird im Folgenden weiter präzisiert: Die (positive) Veränderung von Wissen, Lernprozessen, Erwartungen, Einstellungen und Motivation – die zentrale Aufgabe der Pädagogik.

Dieses Ziel zu erreichen ist nicht einfach. So unterschiedlich die verschiedenen Ansichten der Lehr-Lern-Forschung über die Förderung von Lernprozessen sind (einen guten Überblick findet man beispielsweise bei Helmke 2004), gibt es doch einige Punkte, über die in der Wissenschaft weitgehende Einigkeit besteht. Einer davon lautet: Lernen muss man immer noch selbst (Lerche 2009). Das bedeutet, dass es keine Möglichkeit gibt, von außen Wissen und Können in die Köpfe der Schülerinnen und Schüler zu transportieren, denn der Wissenserwerb ist ein selbstgesteuerter und aktiver Prozess. Die zentrale Leitfrage einer erfolgreichen Förderung lautet daher immer: „Wie hilft man dem Lernenden dabei, sich selbstgesteuert, wohldurchdacht, intensiv und andauernd mit den Inhalten des Unterrichts auseinanderzusetzen? Und was kann die Leistungsbeurteilung dazu beitragen?“

Hierzu ist es notwendig, einen kurzen Blick auf gebräuchliche Lerntheorien zu werfen. Gemäß Piaget (1972, 1981), von Glasersfeld (1992, 1996, 2006) und Collins/Brown/Newman (1989) fasst Lerche (2012) den Prozess des Lernens wie folgt in drei Schritten zusammen:

Der erste Schritt bei der kognitiven Verarbeitung neuer Inhalte ist die Entscheidung, ob der Lernende sich mit den Inhalten weiter beschäftigen will oder nicht. Die dazu notwendige Wahrnehmung ist ein wechselseitiger, lernergesteuerter Prozess zwischen Vorwissen und Information.

Der zweite Schritt im Prozess der kognitiven Verarbeitung ist die Auseinandersetzung der

bedeutsamen Inhalte mit dem Vorwissen, um ein inneres Verständnis für sie zu entwickeln. Dabei wird entschieden, welche Aspekte der neuen Information bereits gut repräsentiert werden (also bereits bekannt sind) und welche neu konstruiert werden müssen. Dieser Schritt birgt die Gefahr einer Übersimplifizierung von Inhalten („Das habe ich schon x-mal gemacht, das ist doch nichts Neues!“) oder dass der Lernprozess abgebrochen wird, da für den Lernenden in der neuen Information keine Ähnlichkeiten mit bereits Bekanntem zu entdecken sind („Das verstehe ich nie!“).

Im dritten Schritt strebt der Lernende an, das kognitive Netzwerk so zu organisieren, dass kein Widerspruch zwischen dem eigenen Wissen und den Informationen aus der Umwelt besteht, um so einen Zustand der Äquilibration zu erreichen (Festinger 1957). Wissensinhalte, die dem Lernenden nicht neu sind, können assimiliert werden (also dem eigenen kognitiven Netzwerk beispielsweise als zusätzlicher Beleg für bekannte Zusammenhänge hinzugefügt werden), wohingegen unbekannte Aspekte akkommodiert werden müssen. Dabei werden in einem Konstruktionsprozess neue Verständnisstrukturen, beispielsweise Differenzierungen bekannter Heuristiken, in einer fortlaufenden Auseinandersetzung zwischen der neuen Information und dem Vorwissen erworben. Genau hier steckt jedoch die zentrale Herausforderung: Lernende wissen (meist intuitiv), dass Akkommodation ein eher anstrengender Prozess ist. Um diese Anstrengung zu umgehen, gibt es eine Reihe von Ausweichmöglichkeiten, die es dem Lernenden auch ohne Konstruktionsprozesse ermöglichen, Äquilibration zu erreichen – beispielsweise Strategien wie „Die Information ist falsch“ oder „Dieses Problem werde ich niemals haben“. Die folgende Abbildung stellt den geschilderten Ablauf grafisch dar. Die Entscheidungspunkte *Bedeutungszumessung, Vorwissenspassung* und *Bereitschaft* bilden die Basis für die Zielerreichung, den Wissenserwerb. Hierbei kann das (diagnostische und unterrichtliche) Handeln der Lehrkraft positiven wie negativen Einfluss auf den Gesamtprozess haben.

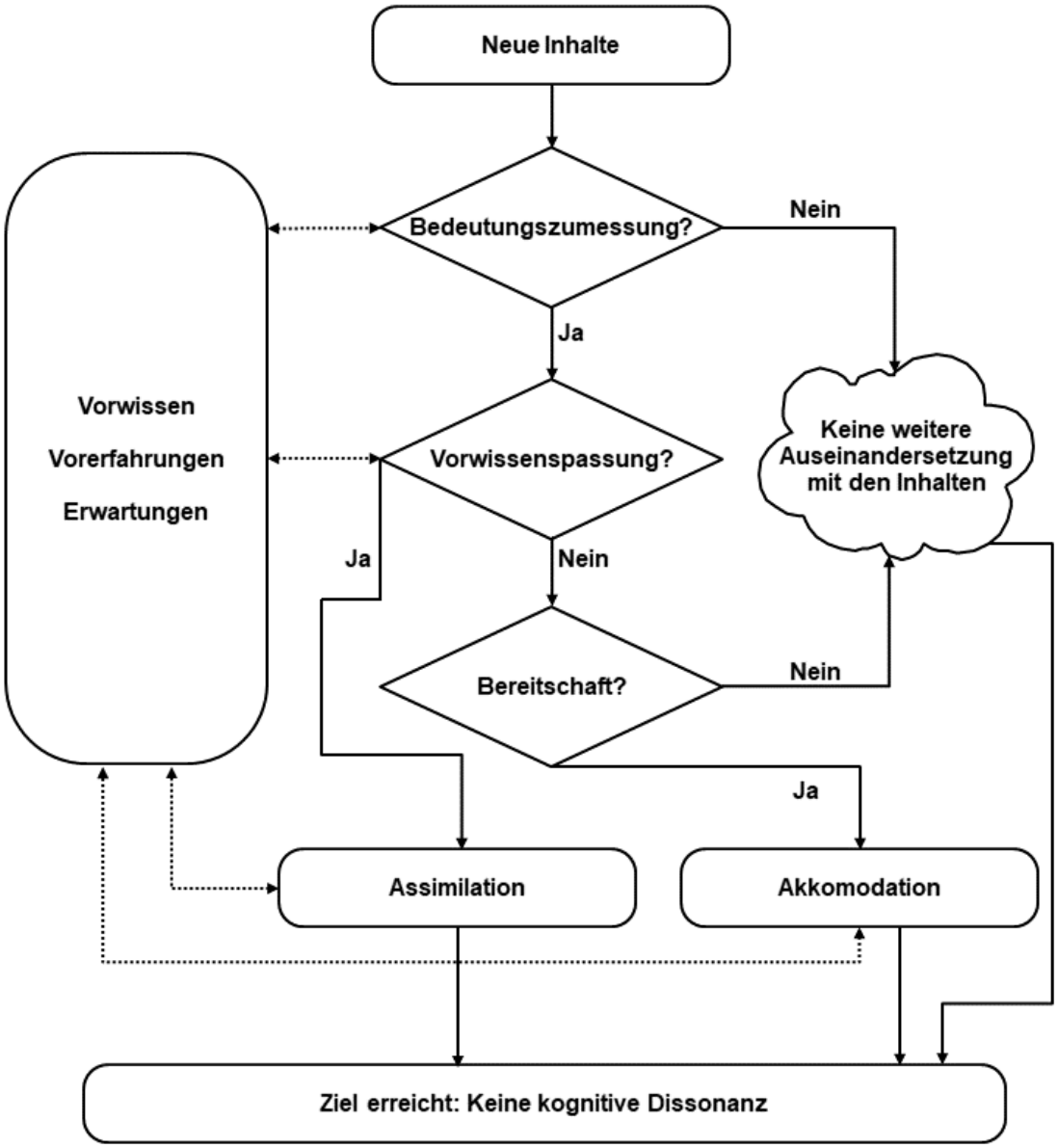

Abb. 5: Modellierung des Lernprozesses, abgeleitet von Piagets Theorie genetischer Erkenntnistheorie

An dieser Stelle muss auf die Bedeutung der Leistungsbeurteilung für diesen Prozess verwiesen werden. Ziel und Aufgabe der Schule ist es, den Lernenden dabei zu unterstützen, dass er Wissen und Fähigkeiten erwirbt, die für eine selbstständige und mündige Gestaltung des eigenen Lebens notwendig und brauchbar sind. Unsere Gesellschaft definiert hierzu eine Reihe von Kompetenzen, die sie für wichtig hält, und beauftragt die Schule als Institution, den Kompetenzerwerb der Kinder zu initiieren. Diese Kompetenzen sind das Ziel. Um ein Ziel zu erreichen, benötigt man immer aufs Neue eine sorgfältige Analyse der Ausgangsbedingungen und des bereits Erreichten (Wo stehe ich?), erst dann kann man einen Weg zum Ziel planen.

Die Analyse dieser Ausgangsbedingungen liefert die schulische Leistungsbeurteilung. Sie ermittelt den Wissensstand und gibt Anregungen, um das Ziel zu erreichen. Dieser Vorgang muss, damit er wirksam ist, kontinuierlich und aktivierend durchgeführt werden.

Genau das ist auch die gemeinsame Botschaft der drei Episoden zu Beginn dieses Kapitels: Es gibt einen gemeinsamen Wunsch von Eltern und Lehrkraft, dass die Kinder etwas lernen. Dass hierbei Konflikte – teils aus Unwissenheit, teils aus falschen Handlungen – entstehen, ist selbstverständlich. Versteht man diese Konflikte aber als konstruktive Kritik, die in erster Linie gegeben wird, weil (begründete oder unbegründete) Zweifel am Fortgang des Lernenden bestehen, können an dieser Stelle Wege gefunden werden, um die Leistungsbeurteilung und damit die Förderung des Lernprozesses zu verbessern.

Sowohl die Instructional-Design-Forschung als auch die Vertreter konstruktivistischen Lehrens und Lernens haben dazu in den letzten Jahrzehnten eine Reihe von Forschungsanstrengungen unternommen. Gut zusammengefasst werden die Ergebnisse dieser Bemühungen von Hattie (2013, 2014), der mit Hilfe einer Untersuchung unterschiedlicher Metaanalysen zum Thema
„schulischer Lernerfolg“ die Wirkung von unterrichtlichen Bemühungen und Methoden auf das Lernen untersucht hat. Er findet eine Reihe sehr interessanter, manchmal auch überraschender Erkenntnisse:

Die Schaffung von erleichternden Bedingungen für den Lernprozess hat kaum positive Auswirkungen, es bringt also nicht viel, wenn man versucht, es dem Lernenden so einfach und angenehm wie möglich zu machen. Methoden und Medien haben für die Aktivierung von Schülerinnen und Schülern ebenfalls keine wesentliche Bedeutung; Lernende werden also nicht dadurch besonders aktiviert, dass sie z. B. in offenen Unterrichtsformen arbeiten dürfen oder die Inhalte mit Filmen und multimedialen Angeboten rezipieren können. Im Ergebnis ist der Einfluss der Lehrkraft auf den Lernerfolg gering, wenn sie versucht, es den Schülerinnen und Schülern möglichst einfach zu machen („facilitator“). Die Wirkung der Lehrperson wird jedoch bedeutsam, wenn sie aktiviert („activator“), also die eigenständige Beschäftigung mit den Inhalten des Unterrichts anstößt und fordert. Als geeignet für die Aktivierung nennt Hattie (ebd.) unter anderem die folgenden Aspekte:

- Handlungsorientiertes Feedback,
- Permanente Selbstreflexion des eigenen Unterrichts,
- Klarheit und Direktheit der Instruktion,
- Herausfordernde Ziele und Lerngelegenheiten,
- Regelmäßige Evaluation des Lernstandes (nicht wegen der Benotung, sondern aus Interesse am Fortschritt der Lernenden).

Für all diese Aspekte ist die Diagnose des Lern-, Motivations- und Einstellungsstandes der einzelnen Lernenden eine unbedingte Notwendigkeit.

Die Prüfung

Der zweite Schritt in der Definition der schulischen Leistungsbeurteilung anhand des Ingenkamp-Modells bildet die Aufforderung zum Leistungshandeln durch den Lehrenden. Dies erfolgt in der Regel durch eine Aufgabenstellung (schriftliche Prüfung) oder eine Frage (mündliche Prüfung), die der Lernende aufnimmt und bewertet. Dann wird er auf der Grundlage seines Verständnisses von der Fragestellung versuchen, eine Passung zwischen Vorverständnis und Aufgabenstellung herzustellen und diese Passung in Handlungen (meist eine schriftliche oder mündliche Antwort) zu überführen.

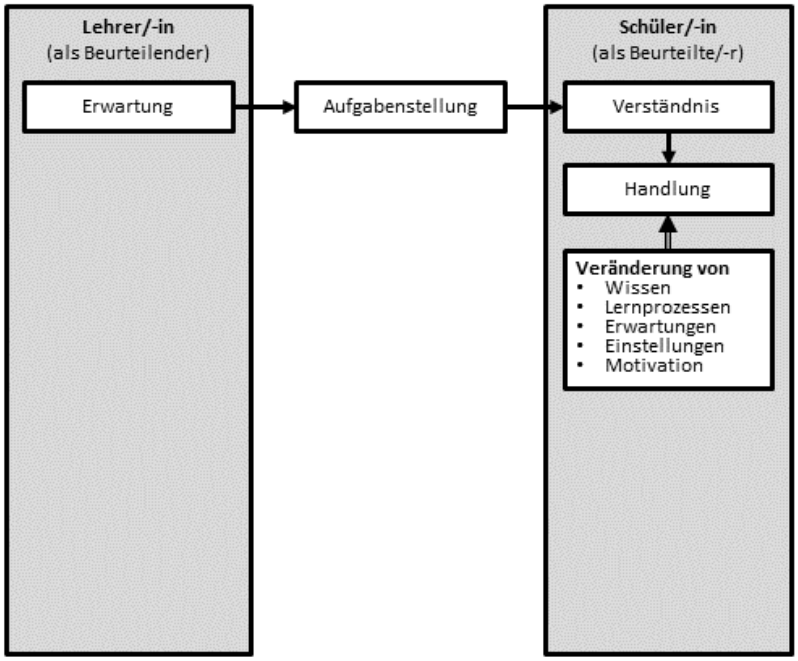

Abb. 6: Modell der pädagogischen Diagnostik (Schritt 2: Die Prüfung)

Die eigentliche Kompetenz bzw. der Lernerfolg sind nicht direkt beobachtbar und damit auch nicht direkt diagnostizierbar. Der Lernende kann seine Kompetenz letztendlich nur anhand durchgeführter Handlungen (zum Beispiel mit Aussagen im Prüfungsgespräch, Referaten, schriftlichen Darstellungen, Berechnungen oder Werkstücken) zeigen, und nur diese Handlungen können Grundlage der Beurteilung des Lernenden sein. Im Kontext der schulischen Leistungsbeurteilung ist vor allem die prüfende, also die ausschnitthafte, auf den Vergleich mit der norm- bzw. funktionsgerechten Ausführung einer Handlung ausgerichtete Beobachtung von Bedeutung (Kiel 2001). Der Beobachtungsausschnitt und das zugrundeliegende Beobachtungsraster werden durch die Prüfung, also durch Aufgabenstellung und erwartete Handlungen, festgelegt und bestimmt (ausführlich dargestellt in Kapitel 7 dieses Bandes). Die Maßstäbe des Bewertungsprozesses generieren sich im schulischen Kontext in der Regel aus den Lehrplänen bzw. den Bildungsstandards (daher spricht man von einer *kriterialen* oder *sachlichen Bezugsnorm*), aus der fachlichen Expertise des Prüfenden, manchmal auch aus dem Vergleich mit früheren Leistungen des Lernenden *(individuelle Bezugsnorm)* – niemals aber aus dem Vergleich des Einzelnen mit den Mitschüler/-innen der Klasse (*soziale Bezugsnorm,* siehe hierzu auch Kapitel 5 dieses Bandes).

Der Prozess der Bewertung

Der eigentliche Bewertungsprozess wird aus den Ergebnissen der Prüfungssituation, also aus den Handlungen des zu Beurteilenden, abgeleitet. Es ergibt sich im Ergebnis der so genannte Handlungsregulationskreislauf (Frese/ Zapf 1994): Die Handlungen des zu Bewertenden werden vom Prüfenden aufgenommen, bewertet und in ein Feedback überführt. Der Lernende wiederum nimmt das Feedback auf, bewertet es und nutzt es im Idealfall, um seine weiteren Lernhandlungen anzupassen und das gemeinsame Ziel, die Veränderung von Wissen und Kompetenzen, zu erreichen. Das folgende Schaubild beschreibt die Handlungsregulation genauer.

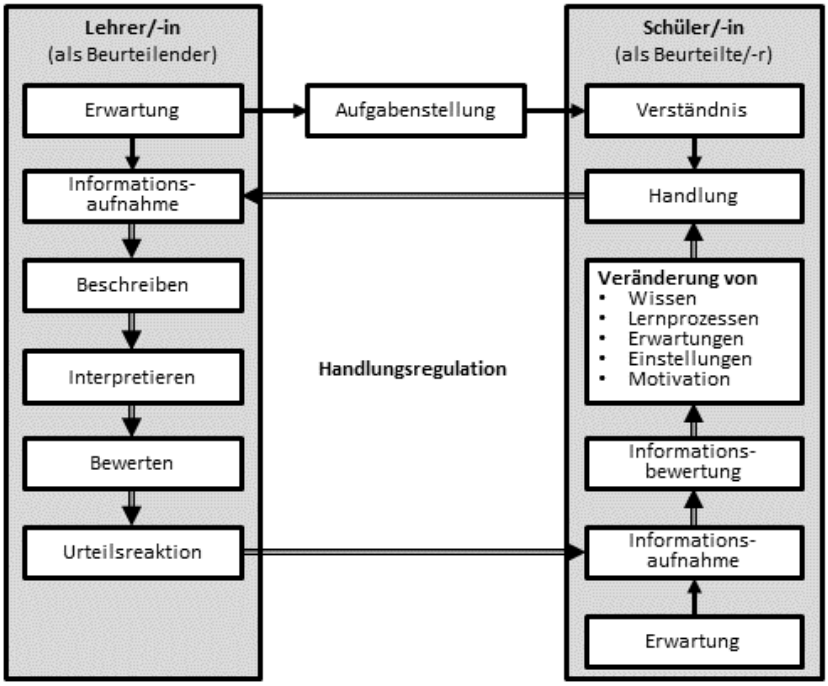

Abb. 7: Modell der pädagogischen Diagnostik (Schritt 3: Der Handlungsregulationskreislauf)

Kiel (2001) empfiehlt, die Bewertung der Handlungen des Prüflings auf die drei Aspekte *Beschreiben, Interpretieren* und *Bewerten* zu stützen. Dadurch sollen eine reflexive, analytische Haltung eingenommen und vorschnelle Beurteilungen möglichst vermieden werden. Diese diagnostischen Prozesse bilden den Kern des Bewertungsprozesses und sind damit auch Bestandteil der weiteren Ausführungen zu diesem Thema. Daher sollen sie an dieser Stelle zunächst definiert und in ihrer Bedeutung für den Gesamtprozess untersucht werden.

Unter einer *Beschreibung* versteht man eine möglichst interpretationsfreie Darstellung der Handlung bzw. des Handlungsergebnisses in einer spezifischen Situation wie beispielsweise einer Prüfung. Der Fokus liegt hierbei auf dem Adjektiv *interpretationsfrei;* es muss also versucht werden, möglichst neutral zu beschreiben, was der zu Bewertende sagt oder tut. Kiel beschreibt diesen Vorgang wie folgt:

„Die Leitfragen auf dieser Ebene lauten: Was für Elemente sind vorhanden? Wie ist die angenommene Beziehung dieser Elemente? Wie wird mit den Elementen und ihren Beziehungen umgegangen? Wie sieht der Bearbeiter die Elemente und ihre Beziehungen? Wie sehen andere Personen möglicherweise diese Elemente und Beziehungen? Hier geht es einerseits um Aspekte wie das Auflisten, Benennen, Definieren und andererseits um das Einwirken auf die Gegenstände des wissenschaftlichen Arbeitens in Form von Experimenten, Befragungen, Herstellen von Textbezügen, Einordnen in Kontexte usw. In beiden Fällen stellen die Bearbeiter möglichst verschiedene Perspektiven dar und fragen nicht nach dem ‚Warum' und ‚Wozu'" (S. 59).

In jedem Falle muss vorschnelles Interpretieren und Bewerten, also die unmittelbare Zuschreibung einer Eigenschaft oder eines Wertes bereits während des Beobachtungsprozesses, vermieden werden, da andernfalls die Kontrolle über diesen Prozess entgleitet (vgl. Kretschmer/Stary 2007). Sonst könnte es z. B. passieren, dass man die Handlungen eines ansonsten eher schlechten Schülers von Anfang an negativer bewertet, als sie es eigentlich sind. Erst eine möglichst genaue Beschreibung der Handlungen lässt eine darauf aufbauende Interpretation und Bewertung der Leistungen zu, sie bildet also eine notwendige Basis des Beurteilungsvorgangs.

Im Kontext der Schülerbeurteilung kann beispielsweise auf der beschreibenden Ebene festgestellt werden, mit welchen Worten die Schülerin / der Schüler die Definition eines Phänomens darstellt, welches Beispiel sie/er für die Illustration des Phänomens gibt, wie sie/er die Bedeutung des Themas für seine

Lebenswirklichkeit einschätzt und wie sich die Beziehung dieser Aussagen zu den fachwissenschaftlich vorgegebenen Sachverhalten darstellt. Beschrieben werden kann aber auch, ob die Antworten flüssig oder stockend und unsicher gegeben werden, ob die Aussagen als direkte Antwort auf die Frage oder als selbstreflexiver Entwicklungsprozess getroffen werden. Für die Förderdiagnostik benötigt man zudem Beschreibungen und Begleitinformationen über den betreffenden Menschen, etwa über Angstverhalten bei Prüfungen oder aktuelle persönliche Umstände.

Der Vorgang des *Interpretierens* fragt nach der Ursache der Handlungen: Warum zeigt der Lernende die Handlungen, die er zeigt? Dieser Vorgang stellt die fachliche und pädagogische Expertise des Prüfers den beschriebenen Handlungen gegenüber und sucht aus diesem Vergleich nach Gründen, warum die Handlungen wie beschrieben durchgeführt wurden. Kiel (ebd.) definiert es als „Schlussfolgern und Zumessen von Bedeutung nach dem Beschreiben durch die Angabe von Gründen für Elemente und ihre Beziehungen" (S. 60).

Im Kontext der oben genannten Schülerbeurteilung kann also beispielsweise aufgrund der Beschreibungen geschlossen werden, ob der Lernende fundiertes Wissen hat und ob er die Bedeutung des Themenbereiches verstanden hat, aber auch, ob eventuell andere Aspekte wie Prüfungsangst, Liebeskummer oder Antipathie eine Rolle für die gezeigten Handlungen spielen. Wichtig dabei ist: Grundlage jeder Interpretation sind Aspekte aus der vorherigen Beschreibung und darauf aufbauende logische Schlussfolgerungen. Natürlich kann (und sollte) das Feedback neben den Sachaspekten auch Dimensionen wie mangelnden Fleiß oder psychische Probleme ansprechen. Es muss solche Aspekte aber immer durch die Beschreibungen und die daraus resultierenden Interpretationen begründen können und darf nicht auf reinen Vermutungen oder Vorurteilen aufbauen.

Unter *Bewertung* versteht man die Einschätzung des Wertes oder der Bedeutung eines Sachverhaltes bzw. Gegenstandes. Sie wird normalerweise anhand eines Bewertungsmaßstabes vorgenommen. Genauer beschreibt es Kiel (ebd.):

„Beim Bewerten geht es darum, über Aspekte, die beim Beschreiben oder Interpretieren genannt wurden, positive oder negative Aussagen zu fällen. […] Die Leitfragen auf dieser Ebene lauten: Welche Elemente und welche Beziehungen sind zu bewerten? Welche Gültigkeit lässt sich durch eine Bewertung herausfordern oder bestätigen? Welcher Maßstab liegt der Bewertung zugrunde? Zu welchen Ergebnissen würden andere mit den gleichen oder mit anderen Maßstäben kommen? Auch hier wird […] – wie bei der interpretierenden Ebene – die Angabe von Gründen erwartet. Solche Bewertungen können einen persönlichen Charakter haben oder standardisierten Verfahren mit klar festgelegten Maßstäben der Evaluierungsforschung folgen (S. 60)."

Bei der Bewertung geht es jedoch neben der reinen Notengebung vor allem um die Begründung und Rückmeldung der Leistungsbewertung. Ihre Gestaltung ist ein bedeutsamer Aspekt zur Erreichung des eingangs erwähnten Ziels, der Anpassung von Lernprozessen, Wissensständen, Einstellungen oder der Motivation. Denn nach wie vor gilt: Ob der Lernende zu dieser anstrengenden Tätigkeit bereit ist, entscheidet er letztendlich selbst. Der Fördergedanke liegt also darin, die Schüler/-innen durch geeignete Rückmeldung ihrer Leistungen in ihren Handlungsintentionen und Handlungsdurchführungen bestmöglich zu unterstützen. Anknüpfend an das Handlungsregulationsmodell von Frese/Zapf (1994) und Hacker (2006) steigt die Wahrscheinlichkeit für Handlungsanpassungen mit der Qualität des Feedbacks. Hierzu muss es

- zeitnah gegeben werden,
- Informationen zur Handlungsausführung transportieren und somit
- in der Lage sein, in Relation zum Lernziel handlungsrelevante Kognitionen zu erzeugen (Nerdinger/Blickle/Schaper 2008).

Damit adressiert ein gutes Feedback die vom Lernenden gezeigten Anstrengungen und Leistungen, nicht dessen Talente und Fähigkeiten, und sollte immer so gegeben werden, dass es möglichst einfach in Handlungen umgesetzt werden kann.

Schwierigkeiten des Beurteilungsprozess

Dieser hier sehr idealisiert beschriebene Bewertungsprozess wird in der schulischen Realität durch eine Reihe von Faktoren beeinflusst, die diesen Prozess ungewiss und damit technisch nicht funktionalisierbar machen. Meist sind es menschliche und damit nicht-triviale bzw. nicht kontrollierbare Aspekte, wie sie im folgenden Schaubild – nicht abschließend – dargestellt werden.

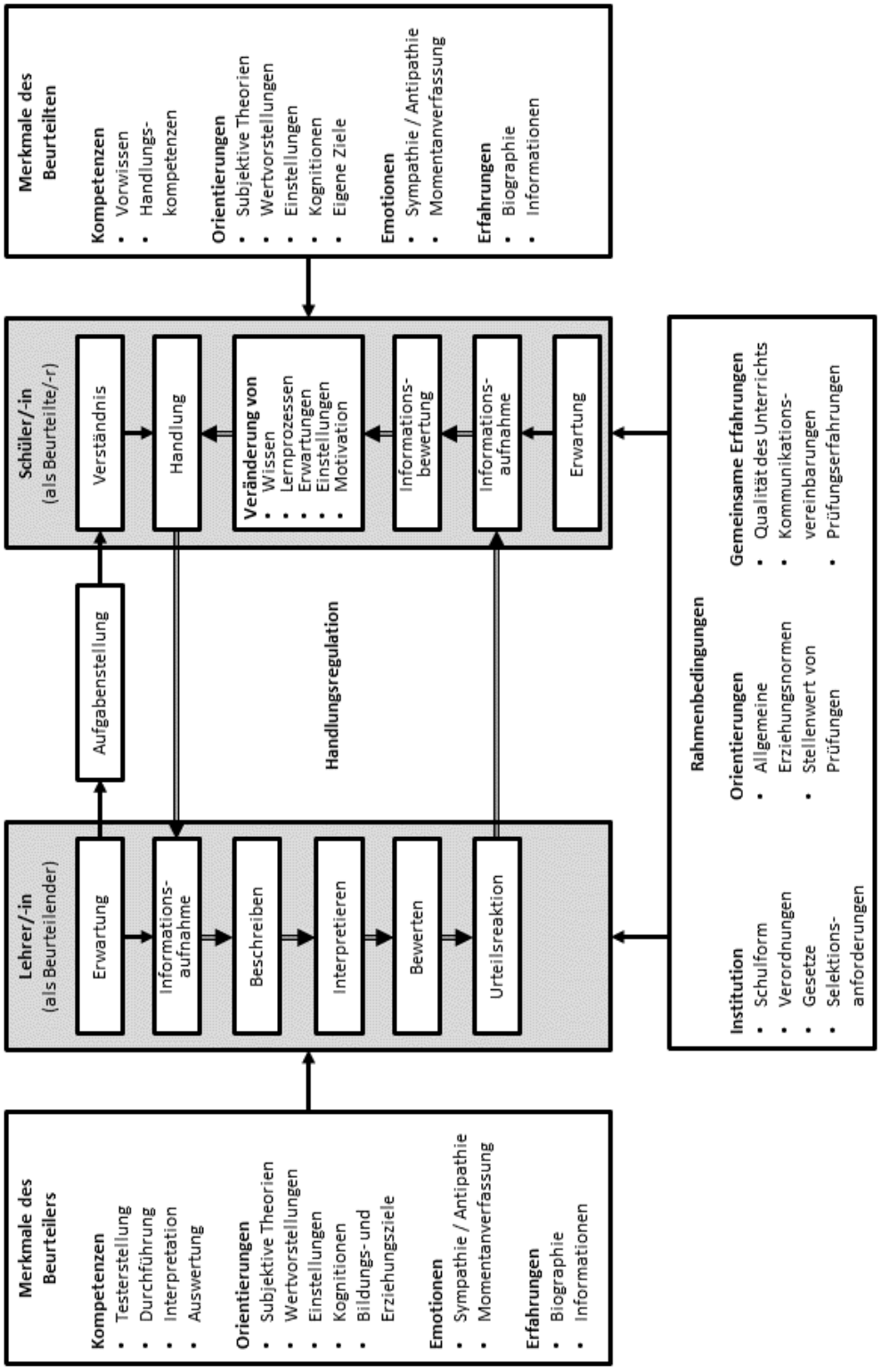

Abb. 8: Modell der pädagogischen Diagnostik (Schritt 4: Einflüsse im Bewertungsprozess)

Das Bemühen, Beobachtungen zu standardisieren und damit die schulische Leistungsbewertung zu professionalisieren, erklärt sich aus einem bekannten psychologischen Phänomen: In einem standardisierten Beurteilungsverfahren ist formal-logisches Beobachten, Denken und Schlussfolgern notwendig. Menschen haben jedoch – besonders unter Zeitdruck – die Tendenz, statt objektiver Regelsysteme vorrangig eigene Heuristiken und subjektive Theorien zu verwenden (Johnson-Laird/Byrne 1991). Dabei können jedoch systematische Fehler auftreten, die beispielsweise aus Vorurteilen, Befangenheiten oder Tendenzen herrühren (Knauff 2005). Wenn Menschen andere Menschen und ihr Handeln beobachten, ist ihre Wahrnehmung geprägt durch langfristig erworbene Wahrnehmungsweisen, Motive und Intentionen:

„Ein Beobachter scheint hauptsächlich diejenigen Ereignisse bevorzugt aufzunehmen, die er am ehesten in sein persönliches Präferenzsystem einordnen kann; dafür lässt er sich am leichtesten sensibilisieren; bei gegenteiligen Ereignissen findet in der Regel demgegenüber eine Wahrnehmungsabwehr statt. Die Wahrnehmung ist praktisch immer das Ergebnis eines Kompromisses zwischen dem, was der Wahrnehmende erwartet, und dem, was objektiv vorliegt" (Schwark 1977, S. 23).

Im Ergebnis kann die Beobachtung durch eine Reihe von Merkmalen beeinflusst werden. Angelehnt an Ingenkamp und Lissmann (2008) können beispielshaft genannt werden:

- *Die Merkmale des Beurteilers.* Hier werden beispielsweise die Beurteilungskompetenzen, subjektive Theorien, Wertvorstellungen, Sympathie, Einstellungen, das Vorwissen, Erfahrungen oder die Momentanverfassung genannt, also zum Beispiel die aktuelle Laune des Beurteilers.
- *Die Merkmale des Beurteilten.* Auch die Erwartungen und Einstellungen des Beurteilten hinsichtlich der Prüfungssituation, der Schule, der Lehrkraft bzw. ihre Kompetenz zum Umgang mit diagnostischen Ergebnissen spielen eine wesentliche Rolle hinsichtlich des Umgangs mit Feedback. Als Beispiel könnte hier das Selbstbewusstsein des Lernenden genannt werden.
- *Rahmenbedingungen.* Zu guter Letzt determinieren auch Aspekte wie die Qualität des vorangegangenen Unterrichts, die spezifische Prüfungssituation, die Prüfungsordnung, allgemeine Erziehungsnormen, Selektionsnormen oder der Stellenwert der aktuellen Prüfung den Beobachtungsprozess. So unterscheiden sich beispielsweise viele Durchführungsbedingungen zwischen den einzelnen Schulformen.

Eine rein objektive Diagnose dieser Handlungen ist daher niemals möglich. Kiel (2007) spricht in diesem Zusammenhang von *kontrollierter Subjektivität* als Ziel einer professionellen Leistungsbeurteilung: In einer komplexen Welt ist eine präzise, objektive Leistungsbeurteilung nicht mehr möglich. Das Ziel der Beurteilung von Schülerleistungen muss es daher sein, die beobachteten Handlungen mit möglichst hoher Genauigkeit und Präzision zu bewerten und zu beurteilen. Dabei handeln Sie unter formal festgelegten Rahmenbedingungen (die meist unstrittig, z.T. sogar gesetzlich vorgegeben sind) und in eigener pädagogischer Verantwortung (was im Kern bedeutet, die Leistungsbewertung sicher, fair, erfolgreich und dauerhaft durchzuführen). Nach Rosemann und Bielski (2001) sind dabei folgende Aspekte wichtig und zu beachten:

- Die Lehrkraft muss sich der grundsätzlichen Subjektivität ihrer Bewertung bewusst sein.
- Sie muss sich bewusst machen, dass es Prozesse interpersonaler Wahrnehmung zwischen Lehrkraft und Schüler/-in gibt und welche Auswirkungen diese Prozesse auf die Leistungsbewertung haben können.
- Sie sollte zwischen der erbrachten Leistung, dem Leistungsverhalten und der Persönlichkeit eines Lernenden trennen und das Feedback immer an der Handlung, nicht an Eigenschaften und Persönlichkeitsmerkmalen ausrichten.
- Sie sollte die Vorteile einer kriteriumsorientierten Leistungsbewertung nutzen (vgl. Kap. 5).

Das definiert auch die Herausforderung der pädagogischen Diagnostik: Wie gelingt es, trotz der vielfältigen einschränkenden Bedingungen den Bewertungsprozess so zu gestalten, dass das Ziel einer qualitativ hochwertigen handlungsregulierenden Rückmeldung möglichst gut erreicht wird?

1.2. Herausforderungen der schulischen Leistungsbeurteilung

Die eingangs vorgestellten fiktiven Episoden aus dem Schulalltag eröffnen eine Reihe von Herausforderungen in Bezug auf die Leistungsbeurteilung. Zur Erinnerung: Von Seiten der Eltern und der Schulleitung können verschiedene Vorwürfe an die Qualität Ihrer Testerstellung und -auswertung vorgebracht werden. Ohne die notwendigen Kompetenzen fiel es Ihnen zu Beginn des Kapitels sicherlich schwer, passende und wohlformulierte Erwiderungen zu finden.

Die anfangs vermutlich wahrgenommene Unsicherheit ist sicherlich auch nach der Lektüre des ersten theoretischen Teils nicht bedeutend geringer geworden, da dieses Kapitel im Kern nur den Rahmen der pädagogischen Diagnostik als Grundlage der Leistungsbewertung an Schulen absteckt. Die Vorwürfe sind jedoch gut nutzbar, um das weitere Vorgehen zu strukturieren: Ausgehend von der grundlegenden Definition der pädagogischen Diagnostik und des Handlungsrahmens der Leistungsbewertung in der Schule werden in den folgenden Kapiteln einzelne Aspekte der genannten Problemstellungen aufgegriffen und tiefergehend bearbeitet. Im Einzelnen sind dies:

Warum geben Sie so strenge Noten? Mein Sohn soll später mal auf eine gute Schule gehen.
Mit diesem Vorwurf wird eine Diskussion über die Auswirkungen der Leistungsbeurteilung, insbesondere im Rahmen der Selektions- und Allokationsfunktion der Schule, angestoßen, die im Kapitel 2 dieses Buches fortgesetzt wird.

Ihr Feedback ist eine Katastrophe!
Hier wird das pädagogische Potenzial der Leistungsbewertung angesprochen und damit der Grundgedanke dieses Kapitels wieder aufgenommen. Kapitel 3 zeigt, welche Möglichkeiten zur Förderung pädagogische Diagnostik bietet.

Warum legen Sie den Punkteschlüssel erst nachträglich fest? Das ist nicht erlaubt!
Dieser Vorwurf adressiert die Frage, was bei der Leistungsbewertung in der Schule ge- und verboten ist. Kapitel 4 definiert den durch die Gesetze und Verordnungen gegebenen Rahmen, innerhalb dessen die Beurteilung schulischer Leistungen stattfindet.

Mein Sohn hat sich doch so verbessert. Warum bekommt er immer noch die schlechteste Note?
Hier wird der Bezugsmaßstab der Beurteilung angesprochen. Kapitel 5 stellt die Frage, unter welchen Bedingungen welche Maßstäbe bei der Leistungsbewertung anzusetzen sind.

Die Klausur war zu schwer! Die Fragen waren nicht klar! Sie haben gar nicht durchgenommen, was Sie in der Klausur gefragt haben!
Diese Vorwürfe stoßen die Frage nach der Qualität der Leistungsprüfung an. Hier sind in erster Linie die Kriterien für einen guten Test zu nennen, wie sie im Kapitel 6 besprochen werden.

Wie kommen Sie denn eigentlich zu Ihrem Urteil? Ich habe den Eindruck: Was Ihnen gefällt, bekommt gute Noten, was Ihnen nicht gefällt, bekommt schlechte Noten!
Ein wichtiger Aspekt der Leistungsbewertung ist der bereits angesprochene Weg von der Handlungsbeobachtung zur Urteilsbildung. Auf diesem Weg gibt es, wie oben gezeigt, eine Reihe von Fehlerquellen, Fallstricken und Irrtümer, die durch eine wohldurchdachte Beschreibung, Interpretation und Bewertung minimiert werden können. Kapitel 7 zeigt, wie dies gelingt.

Wie kommen Sie darauf, bei einer Gruppenarbeit jedem die gleiche Note zu geben. Mein Patrick hat alles alleine gemacht, die anderen sollen schlechter bewertet werden!
Kapitel 8 gibt zum Schluss dieses Buches einen Ausblick auf die Leistungsbeurteilung bei offenen Lernformen.

1.3. Bilanzierung des Gelernten

In diesem Kapitel werden folgende Lernimpulse gesetzt:

- Lehrersein ist eine Profession. Um Leistungen angemessen und belastbar bewerten zu können, müssen Sie sich professionalisieren. Dazu gehören laut Herbart sowohl praktische Erfahrung als auch theoretisches Wissen.
- Das Ingenkamp-Modell ermöglicht eine Einordnung von Beurteilungsprozessen in den Kontext von Lehrerhandeln und Lehr-Lern-Prozessen.
- Beurteilungen bilden einen Handlungsregulationskreislauf, in dem die Lehrkraft die Handlungen des Lernenden beschreibt, interpretiert und bewertet. Daraufhin gibt sie ihm durch eigene Handlungen Rückmeldungen, die wiederum vom Lernenden interpretiert und umgesetzt werden und so zu einer Veränderung des Lernhandelns führen können.
- Der eigentliche Lernerfolg einer Person ist nicht sichtbar: Ein Lernender kann seinen Lernerfolg nur durch Handlungen, etwa durch Antworten auf Fragen zeigen, und nur diese können beurteilt werden.
- Da zusätzlich noch Merkmale des Beurteilenden, des Beurteilten und der Beurteilungssituation die Bewertung beeinflussen können, ist eine absolut objektive Beurteilung im schulischen Rahmen nicht möglich. Das Ziel muss es daher sein, zu einer kontrollierten Subjektivität zu gelangen, die sich ihrer Möglichkeiten, aber auch ihrer Grenzen bewusst ist.

Bevor Sie im Text weiterlesen, empfehlen wir Ihnen, die zu Beginn des Kapitels durchgeführte Reflexionsschleife an dieser Stelle – unter Verwendung der neuen Informationen – zu wiederholen.

2. Funktionen der Leistungsbeurteilung

Vielerorts ranken sich gesellschaftliche Mythen über das Vorgehen bei der Leistungsbeurteilung an Schulen. Meist münden diese Vorstellungen in eine generelle Diskussion um die Abschaffung von Ziffernnoten. Die Kernargumente hierfür werden häufig in der Funktion der Schule selbst verortet: Sie solle die Schülerinnen und Schüler zu mündigen, selbstständigen und emanzipierten Menschen erziehen, was jedoch nicht mit einer schablonenhaften Einordnung in fünf oder sechs Notenstufen in Einklang zu bringen sei. Hinzu kommen noch die Infragestellung der Aussagekraft von Ziffernnoten, die nie ganz zu vermeidenden systematischen Fehler bei der Zensurengebung, das klasseninterne Bezugssystem, die Qualitätsunterschiede in den einzelnen Schulfächern und der mangelnde prognostische Wert von Schulnoten für den weiteren Lebensweg.
Illustriert werden kann diese Kritik mit den folgenden Episoden:

Episode 1: Kritik am Punkteschema
Die Eltern Ihrer Schüler/-innen zeigen sich unzufrieden mit dem Notenschema der letzten Mathematikprobe. Die Kritik äußert sich insbesondere darin, dass verschiedene Aufgaben und Fragetypen die gleiche Anzahl an Punkten erhalten. Die geforderten Kompetenzen seien doch gar nicht miteinander vergleichbar; es sei daher nicht verständlich, warum es für Rechenleistungen genau so viele Punkte gebe wie für die Sauberkeit und Sorgfalt einer Konstruktion. Ihr Punkteschema wird als willkürlich klassifiziert und eine entsprechende Beschwerde angedroht.

Episode 2: Bewertung im Religionsunterricht
Als Lehrer/-in der evangelischen Religionslehre wünschen Sie sich bessere Möglichkeiten der Leistungsbeurteilung. Die im Lehrplan festgesetzten Kompetenzziele, beispielsweise „eigene Fragen nach Gott und Welt stellen und offen halten" oder „Symbole und Erfahrungen von Menschen miteinander in Beziehung bringen" lassen sich nur sehr schwer in Notenstufen überführen. Sie beginnen sich zu fragen, ob in solchen – oder vielleicht sogar in allen – Fächern das bekannte Notensystem nicht besser durch Verbalbeurteilungen ersetzt werden sollte.

Episode 3: Das Verbalzeugnis wird unterschiedlich verstanden
Immer wieder kommen Eltern in Ihre Sprechstunde und bitten um eine Interpretation der Verbalbeurteilungen in den Grundschulzeugnissen ihrer Kinder. Die Eltern sind unsicher, wie sie Formulierungen wie die folgenden verstehen sollen:

- *Du hattest keine Schwierigkeiten, dich in den Schulalltag einzugewöhnen, und du hast rasch ein gutes Verhältnis zu deinen Mitschülern gefunden. Du bemühst Dich sehr, den Anforderungen des Unterrichts zu folgen.*
- *Am Unterrichtsgeschehen beteiligst du dich eifrig und wissbegierig und lieferst wohl überlegte Beiträge. Bei der Anfertigung deiner schriftlichen Aufgaben arbeitest du selbständig und richtig, jedoch noch etwas zu langsam.*
- *Du hast den Leselehrgang sehr erfolgreich abgeschlossen. Du kannst Texte fließend, selbständig und sinnerfassend vortragen. Es fällt dir nicht schwer, den erlernten Wortschatz sprachlich richtig anzuwenden und Beobachtungen und Erlebnisse lebendig darzustellen. Texte kannst du zwar langsam, aber sauber und fehlerfrei abschreiben.*
- *Du bekundest Interesse und Freude am Musikunterricht. Dein Flötenspiel könntest du mit etwas mehr Geduld und Freude am Üben weiter verbessern.*
- *Im Kunstunterricht zeigst du viel Fantasie und ein sicheres Form- und Farbgefühl. Der Sportunterricht bereitet dir viel Freude und deine Anstrengungsbereitschaft ist lobenswert.*

Immerhin sind Bewertungen wie „Du bemühst Dich sehr …" oder „könntest Du weiter verbessern" in Arbeitszeugnissen meist als vernichtende Kritik zu sehen. Auffallend ist zudem, dass fast alle Eltern nach dem Notenpendant für die jeweiligen Formulierungen fragen, denn so könnten sie die Leistungen ihres Kindes im Hinblick auf zukünftige Anforderungen besser einordnen.

Die Diskussionen über die schulische Leistungsbeurteilung entzünden sich in den meisten Fällen an den Ziffernnoten. Dabei ist es problematisch, dass Fachwissen und Überblick in diesem Themengebiet sehr einseitig verteilt sind: Lehrerinnen und Lehrer haben ebenso wie Bildungswissenschaftler einen deutlich professionelleren Bezug zu diesem Thema als Eltern und Schüler/-innen, deren Argumente meist von der eigenen Befindlichkeit determiniert werden. Schwierig wird es immer dann, wenn aus dieser Befindlichkeit heraus Generalisierungen und pauschale Kritik an der Zensurengebung abgeleitet werden und darauf aufbauend die Qualität der schulischen Leistungsbeurteilung als solche in Frage gestellt wird.

Um diese Diskussion führen zu können, müssen zunächst die Funktionen der Ziffernnoten geklärt werden.

2.1. Funktionen der Ziffernnoten

Die bekannteste Aufzählung der verschiedenen Funktionen von Schulnoten stammt von Zielinski (1975). Er nennt zehn – oft widersprüchliche – Funktionen, hier zitiert nach Körber 2001:

- Die Rückmeldefunktion für die Lehrerin und den Lehrer: An Hand der Zensurenverteilung soll die Lehrkraft den Erfolg ihres Unterrichts ablesen können.
- Die Rückmeldefunktion für die Schülerinnen und Schüler: Die Note informiert sie über ihren Leistungsstand im Vergleich zu den Mitschülerinnen und Mitschülern.
- Die Berichtfunktion: Durch Noten werden die Eltern über den Leistungsstand ihrer Kinder informiert.
- Die Anreizfunktion: Zensuren dienen zur Motivation der Schülerinnen und Schüler, um sich mit dem dargebotenen Lernstoff zu beschäftigen.
- Die Disziplinierungsfunktion: Durch schlechte Noten werden leistungsunwillige Schülerinnen und Schüler bestraft, um sie dadurch dem gewünschten Leistungsverhalten näher zu bringen.
- Die Sozialisationsfunktion: Schülerinnen und Schüler müssen sich mit Leistungsnormen auseinandersetzen, die sich teilweise von in der Familie gültigen stark unterscheiden. Vor allem nehmen die Schülerinnen und Schüler wahr, so Zielinski, „dass es als fair gilt, wenn unterschiedliche Leistungen auch unterschiedlich belohnt werden."
- Die Klassifizierungsfunktion: Durch die Unterschiede in den Noten werden die Schülerinnen und Schüler in unterschiedliche Bewertungsklassen eingeteilt. Diese Maßnahme ist Grundlage für Förderungsund Selektionsmaßnahmen.
- Die Selektionsfunktion: Besonders gute bzw. schlechte Schülerinnen und Schüler sollen mit Hilfe von Zensuren ausgewählt werden, um sie entsprechenden Institutionen zuführen zu können.
- Die Zuteilungfunktion: Mit der Zensierung werden Berechtigungen für den weiteren sozialen Aufstieg vergeben oder verwehrt.
- Die Chancenausgleichsfunktion: Besonders benachteiligte Schülerinnen und Schüler erhalten bessere Noten, „als es die objektiven Leistungen rechtfertigen würden."

Die Punkte 1 bis 3 sowie 7 bis 9 sprechen die Kernfunktion von Ziffernnoten an: Sie geben allen Beteiligten Rückmeldung und ermöglichen darüber hinaus für den einzelnen Lernenden berufliche und schulische Wege – was streng genommen ebenfalls als Rückmeldefunktion gesehen werden muss, in diesem Falle an die weiterführende Schule, die Universität oder den Arbeitgeber. Die Punkte 4–6 müssen hingegen höchst kritisch gesehen werden. Zu nennen seien hier beispielhaft die Arbeiten von Cohen (1994), Deci und Ryan (1993), Prenzel (1997) oder Hattie (2013, 2014), nach denen intrinsische Motivation durch extrinsische Beurteilung verringert werden kann. Will man beispielsweise bei den Kindern das Interesse an Büchern wecken, so erreicht man durch inhaltliche Prüfungen von Fakten und Verständnis oft genau das Gegenteil: Bücher werden als unheilvolles Gut angesehen, das ursächlich zu Stress und schlechten Noten führt. Leistungsbeurteilung und Lernziel sind in diesem Beispiel nur schwer vereinbare Gegensätze. Gleiches gilt

für die Disziplinierungsfunktion, die zumindest zweifelbehaftet ist (Wengert 2000). Auch die Sozialisationsfunktion der Noten, also die Nutzbarmachung von Informationen der Leistungsrückmeldung zur Unterstützung der eigenen gesellschaftsbezogenen Entwicklung, wird selten als positiver Effekt dargestellt (Jachmann 2003). Allerdings ist auch hier fraglich, ob die Note selbst die Zielscheibe der Kritik ist oder die Beurteilung der Schülerleistungen schlechthin. Es darf durchaus die Frage gestellt werden, ob die Kritik hier nicht an der falschen Stelle ansetzt. Dies soll im Folgenden weiter ausgeführt werden.

2.2. Kritikpunkte an schulischer Leistungsbeurteilung

Jürgen Oelkers (2002) fasst unter Bezugnahme auf Ingenkamp (1971) einige der oft genannten Kritikpunkte an der schulischen Leistungsbeurteilung zusammen:

- subjektive Fehlerquellen,
- das klasseninterne Bezugssystem,
- die Qualitätsunterschiede in den einzelnen Schulfächern und
- den prognostischen Wert.

Subjektive Fehlerquellen. Bei den subjektiven Fehlerquellen geht es um individuelle und unsystematische Fehler, die während des gesamten Beurteilungsvorgangs auftreten können. Die bekanntesten dieser Fehlerquellen werden im Folgenden vorgestellt:

Implizite Wahrnehmungsfehler. Bekannt sind solche logischen Fehler unter Namen wie *Pygmalion-Effekt* (Einbeziehung impliziter Vorannahmen in die Bewertung), *Halo-Effekt* (Einbeziehung besonders hervorstechender Merkmale oder Leistungen eines Lernenden in die Bewertung) oder *soziale Stereotypisierung* (generelle Auf- oder Abwertung eines Geschlechts, von Schüler/-innen mit Migrationshintergrund, mit Behinderung oder von Angehörigen einer Jugendkultur). Diese Fehler werden immer dann wirksam, wenn persönliche oder biographisch bedingte Einstellungen gegenüber einzelnen Schüler/-innen oder Personengruppen, die auf Grund von früheren Erfahrungen (vermeintlich) stabil gewonnen wurden, einen Einfluss auf die Leistungsbemessung haben. Prominentestes Beispiel hierfür sind die Studien von Rosenthal und Jacobson zum oben erwähnten Pygmalion-Effekt (1966). Lehrenden wurde eine Versuchsklasse übergeben, deren Fähigkeiten vorab bestimmt und damit bekannt waren. Dem einzelnen Lehrenden wurden jedoch bei 20 % der Lernenden zufällige, fiktive Vorabbewertungen genannt. Es konnte gezeigt werden, dass die Bewertungen der Lehrkraft stark von diesen fiktiven Vorabinformationen beeinflusst waren. Diese Befunde konnten 1974 von Chaiken, Sigler und Derlega repliziert werden.

Unzureichende Bezugsrahmen. Am häufigsten zitiert wird die (unangemessene und gesetzlich nicht erlaubte) *Verwendung der sozialen Bezugsnorm.* Dabei wird die Leistung eines einzelnen Kindes mit den Leistungen einer sozialen Gruppe, zum Beispiel der Klasse, vergleichen. Die Folge ist, dass die besten Schüler immer sehr gute, die schlechtesten immer ungenügende Ergebnisse erhalten, ganz egal wie gut ihre Leistung wirklich war (vgl. Kapitel 5). Weniger bekannt sind *Reihungsfehler,* die auftreten, wenn bei der Korrektur schriftlicher Arbeiten die vorangegangenen Korrekturen einen bedeutsamen Einfluss auf die implizite Qualitätseinschätzung der aktuellen Arbeit haben. Wenn man beispielsweise mehrere sehr schlechte Arbeiten hintereinander korrigiert hat, könnte bei der Korrektur einer weiteren, vergleichbar mangelhaften Klausur ein milderer Maßstab angelegt werden. Auch die *Kontrast- oder Ähnlichkeitsfehler* fallen in diese Kategorie. Hier vergleicht der Bewertende unbewusst die eigenen Fähigkeiten mit denen der Schüler/-innen, anstatt ein standardisiertes Bewertungskriterium anzulegen. Um es anschaulich zu machen: Durch die fortwährende Beschäftigung mit einem Thema erlangt eine Lehrkraft in der Regel mit den Jahren eine sehr hohe Expertise in ihrem Fach. Der Kontrast zwischen seinen eigenen Kompetenzen und denen der Lernenden wächst stetig. Berücksichtigt der Beurteilende diese Erkenntnis in seinen Bewertungshandlungen nicht, sind Aussagen wie „Die Schüler werden immer dümmer" schnell getroffen.

Tendenzen. In Abgrenzung zu den impliziten Persönlichkeitstheorien, bei denen einzelne Lernende beispielsweise aufgrund von Sympathie milder beurteilt werden, betreffen Tendenzen alle zu Beurteilenden. Bekannt sind hier die *Tendenz zur Strenge* und die *Tendenz zur Milde,* bei denen die Bewertungen

grundsätzlich eher im oberen oder im unteren Bereich angesiedelt werden. Erklärt werden kann dies mit unterschiedlichen Erwartungshorizonten. So setzt beispielsweise eine strenge Lehrkraft den Erwartungshorizont sehr hoch an (im beinahe fehlerfreien Bereich) und reduziert die Note mit der Anzahl der auftretenden Fehler. Ebenso fehlerhaft ist eine generelle *Tendenz zur Mitte,* bei der sich die überwiegende Mehrheit der vergebenen Zensuren im mittleren Bereich der Notenskala bewegt. Die Gründe hierfür liegen in der Gestaltung der Tests (eine hohe Anzahl von Aufgaben mit mittlerem Schwierigkeitsniveau), in der Verwendung eines zu umfangreichen Kriterienkatalogs (wenn sehr viele Aspekte bewertet und gemittelt werden, ergibt sich die Tendenz zur Mitte auf arithmetischem Wege) oder in der Furcht des Beurteilenden, stark nach oben und unten abweichende Beurteilungen inhaltlich begründen zu müssen.

Individuelle Bevorzugungen oder Benachteiligungen. In dieser Kategorie werden Unkorrektheiten angesprochen, die individueller Natur sind, die also die Auf- oder Abwertung der Leistungen eines bestimmten Lernenden betreffen. Der bekannteste davon ist der *Wissen-um-die-Folgen-Fehler.* Ein Urteil fällt milder aus, wenn die beurteilende Lehrkraft weiß, dass daraus in der Zukunft negative Konsequenzen resultieren, wenn also beispielsweise eine Note verbessert wird, um das Sitzenbleiben eines Lernenden zu vermeiden. Dies kann durchaus pädagogisch motiviert sein, ob in diesem Fall von einem Fehler gesprochen werden kann, sei deshalb dahingestellt. Definitiv falsch wäre es hingegen, einem Lernenden, wie im Eingangsbeispiel zu Kapitel 1 von den Eltern gefordert, im Zweifelsfall eine bessere Note zu geben, nur weil er später einmal eine gute Schule besuchen soll. Auch der *Sympathie-Effekt,* also die Berücksichtigung der persönlichen Vorlieben und Antipathien für einzelne Lernende bei der Bewertung und der *Status-Quo-Effekt,* der die Tendenz der Lehrkraft beschreibt, trotz einer Veränderung der objektiven Leistungen am Status-Quo des Lernenden festzuhalten, fallen in diese Kategorie.

Grundsätzlich beeinträchtigen alle genannten Fehler den Bewertungsprozess und machen ihn unscharf. Es ist allerdings zu kurz gegriffen, daraus die generelle Unbrauchbarkeit schulischer Leistungsbeurteilung zu folgern. Die genannten Fehlerquellen können verringert werden, wenn einige allgemeine Hinweise beachtet werden (Jürgens 2005, S. 150 ff.):

- Die Vermeidung von Fehlern bei der Leistungsbewertung setzt das Wissen um mögliche Fehlerquellen voraus. Es schärft das Bewusstsein für die besonderen Herausforderungen im Bewertungsprozess.
- An die Kenntnis der Fehlerquellen muss regelmäßig eine kritische Reflexion der eigenen Bewertungspraxis anschließen.
- Die diagnostische Kompetenz des Kollegiums ist eine hervorragende Ressource für diese Analyse.
- Dabei sollte das Bewusstsein, dass eine Beurteilung einer Leistung auf Grund einer systematischen Beschreibung, Interpretation und Bewertung von konkreten Handlungen entsteht, maßgebend für den Beurteilungs- und den Reflexionsprozess sein.
- Eine besondere Aufmerksamkeit ist auf die Interpretation der Beobachtungen zu legen, um das Beurteilungsergebnis sachlich und nachvollziehbar zu begründen.

Diese Handlungsempfehlungen sind seit langem Inhalt der Lehramtsausbildung und führen in der Regel zu dem erwünschten Ergebnis. Es hat sich in den letzten Jahren – verstärkt durch die Erkenntnisse der aktuellen Standardisierungsbemühungen – vieles zum Positiven verändert; eine generelle Kritik an der schulischen Leistungsbewertung, lediglich basierend auf dem Vorhandensein der genannten subjektiven Fehler, ist nicht mehr angebracht.

Das soziale Bezugssystem

Der zweite Vorwurf Oelkers (2002) zielt darauf ab, dass Lehrerinnen und Lehrer die Tendenz haben, Noten auf Grund eines klasseninternen Bezugsrahmens zu vergeben. Das hat mehrere unerwünschte Effekte, die in Kapitel 5 dieses Buches noch weiter ausgeführt werden:

- Wenn man das Pech hat, in einer guten Klasse zu sein, wird die eigene Leistung negativer bewertet.

- Beurteilungen auf Basis der sozialen Bezugsnorm bilden damit möglicherweise eine ungünstigere Ausgangsposition für die weitere Bildungslaufbahn.
- Da immer nur die Klassenbesten gute Noten bekommen, muss man andere Schüler/-innen leistungsmäßig überholen, um besser zu werden. So kommt es innerhalb der Klasse zu einer Konkurrenzsituation.
- Der eigentliche Lernfortschritt wird ausgeblendet, da ja die ganze Klasse besser wird und sich die eigene Note immer an der Position in der Klasse und nicht am zusätzlichen Wissen misst.
- Gerade bei schwächeren Schüler/-innen ist die Verwendung der sozialen Bezugsnorm motivational ungünstig.

Zum letzten Punkt zitiert Oelkers zwei Studien (Huberman 1980; Rheinberg 1980), die einen Nachweis erbringen, dass die soziale Bezugsnorm sich ungünstig auf Leistungsmotivation, Anstrengungsbereitschaft, Ursachenzuschreibung von Schulleistungen sowie die Selbstwahrnehmung auswirkt.

Die Nachteile und unerwünschten Effekte der sozialen Bezugsnorm sind jedoch seit langem bekannt, sodass ihre Verwendung bei der schulischen Leistungsbeurteilung per Gesetz ausgeschlossen ist. Gefordert wird vielmehr die Verwendung transparenter Kriterien. Auch hier hat sich in den letzten Jahrzehnten vieles zum Positiven verändert, wenngleich die soziale Bezugsnorm an manchen Schulen immer noch vorkommt.

Die Varianz der Noten zwischen den verschiedenen Fächern

Mit diesem Vorwurf wird angesprochen, dass die Notenverteilungen zwischen den einzelnen Fächern höchst unterschiedlich sind, dass also in den verschiedenen Fächern mit unterschiedlicher Strenge bewertet wird. Tatsächlich zeigen entsprechende Untersuchungen, dass umso strenger beurteilt wird,

- je häufiger die Leistungen in schriftlichen Arbeiten überprüft werden,
- je leichter die Leistungen quantifizierbar sind
- und/oder je stärker die sprachlichen Anforderungen einer Prüfung hervortreten (Ziegenspeck 1999, S. 137 ff.).

Kornadt (1978) untersuchte diesen Vorwurf im Rahmen des Schulversuchs *Oberstufe Saar* genauer. Die Ergebnisse stützen seine Vermutung: Es gibt, wie die nachfolgende Grafik zeigt, große Unterschiede zwischen den einzelnen Fächern.

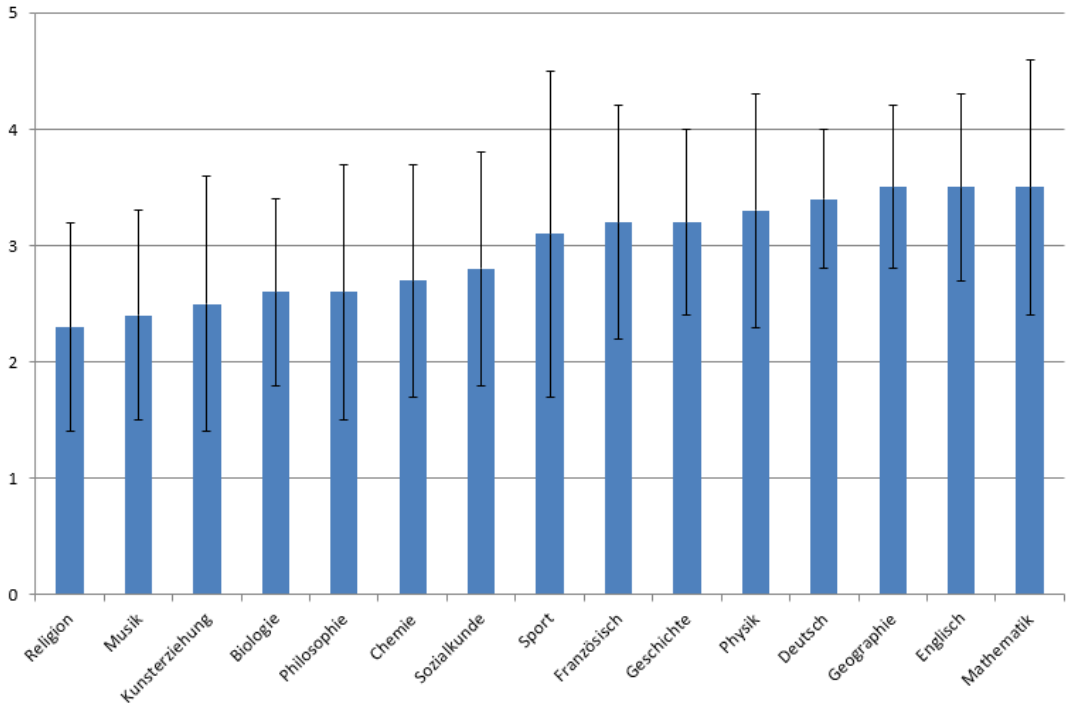

Abb. 11: Mittelwerte und Standardabweichungen der Durchschnittsnoten verschiedener Fächer. Durchschnittswerte aus mindestens 4 Klassenarbeiten. N pro Fach: 80–355. Quelle: Kornadt, H.-J. (1978). Abschlussbericht über die wissenschaftliche Begleituntersuchung zum Schulversuch Oberstufe Saar 1970–1977. S. 283.

Dabei sind bei den einzelnen Fächern nicht nur Unterschiede im Durchschnitt der Noten, sondern auch Unterschiede in der durchschnittlichen Notenverteilung zu erkennen:

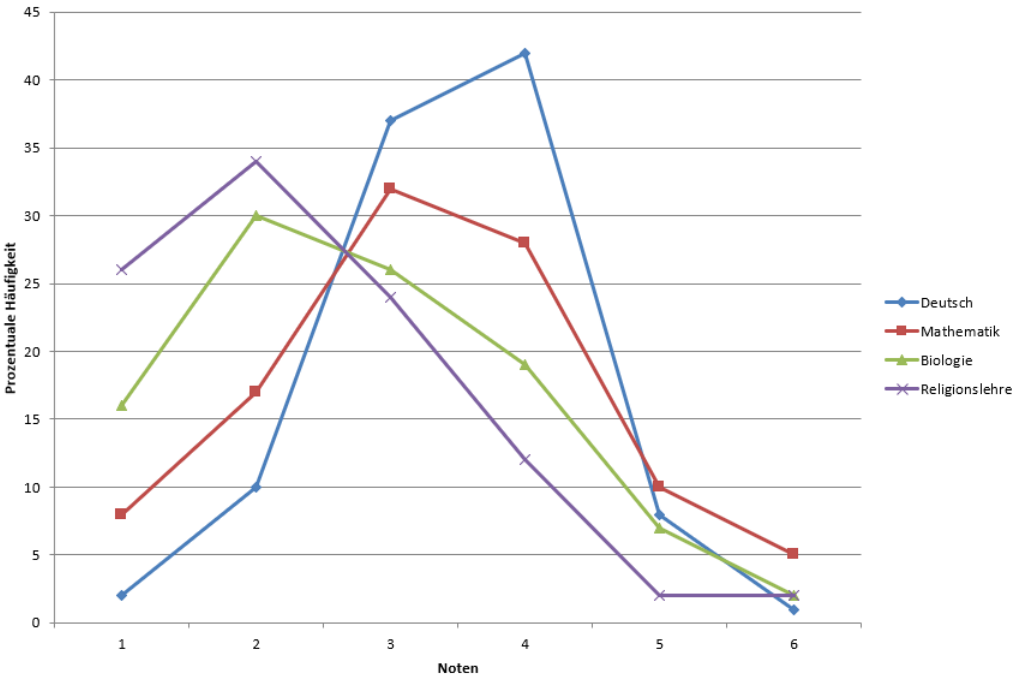

Abb. 12: Notenverteilung in einzelnen Fächern. Quelle: Kornadt, H.-J. (1978). Abschlussbericht über die wissenschaftliche Begleituntersuchung zum Schulversuch Oberstufe Saar 1970–1977. S. 288.

Ob dieses Ergebnis jedoch auf die Unbrauchbarkeit schulischen Beurteilens schließen lässt, darf bezweifelt werden. Schließlich gilt, dass die gemessenen Unterschiede in der Strenge bzw. in den Durchschnittsnoten und den Notenverteilungen jeweils die gesamte Kohorte betreffen. Das heißt: Wenn alle Schülerinnen und Schüler die gleichen Fächer besuchen, ist trotz der Unterschiede die Vergleichbarkeit der Noten gegeben. Dies gilt mittlerweile auch für die gymnasiale Oberstufe. War es in früheren Jahren durch die relativ große Wahlfreiheit zwischen den einzelnen Kursen noch möglich, die eigene Abiturnote durch die Wahl vermeintlich leichter und milde benoteter Kurse positiv zu beeinflussen (in Bayern war der Klassiker hierfür die Leistungskurskombination Latein/Kunst), ist diese Möglichkeit durch den hohen Grad an verpflichtenden Fächern (Mathematik, Deutsch, eine Sprache, eine Naturwissenschaft) inzwischen keine signifikante Einflussgröße mehr. Der – durchaus reale – Vorwurf der unterschiedlichen Strenge zwischen den einzelnen Fächern trägt damit nicht bedeutsam zur Unschärfe des Notensystems bei.

Die mangelhafte Prognosefähigkeit von Schulnoten

Auch diese Wahrnehmung ist auf Ingenkamp (1971) zurückzuführen. Im Kern lautet der Vorwurf, dass Noten zwar den Erfolg eines Lernenden in der nächsthöheren Klassenstufe recht gut prognostizieren, sich allerdings für den weiteren Schul- oder Studienerfolg keine verlässlichen Aussagen treffen lassen.

Jakobs (2010) liefert für diese Behauptungen konkrete Zahlen. Demnach liegen die Korrelationen[1] zwischen den Vornoten (Klassenarbeiten der Lehrkraft, nur für ihre Klasse) und der zentralen Abschlussklausur (mit gleichen Anforderungen für Schüler/-innen aller Klassen) im Bereich zwischen 0,36 (für das Fach Deutsch) und 0,70 (für das Fach Mathematik). Die gemittelte Prognosefähigkeit von Abschlussnoten (Abiturdurchschnitt) für den Studienerfolg (Examensnote) wurde ebenfalls ermittelt, sie liegt bei 0,46.

Wenn man die Tendenz zur guten Note und dadurch bedingt die geringe Streuung der Abschlussexamina an der Universität in Betracht zieht, dann sind die prognostischen Qualitäten der Abiturdurchschnittsnote, trotz ihrer augenscheinlich geringen Höhe von 0,46, erstaunlich gut. Sie liegen damit höher als die aller anderen messbaren Einflussgrößen, z. B. die der Lehrqualität, des Curriculums sowie die der gesellschaftlichen, individuellen oder universitären Rahmenbedingungen.

Die Prognosefähigkeit des Abiturnotendurchschnitts für den Berufserfolg (u. a. Berufsposition, Einkommen usw.) fällt geringer aus und liegt zwischen 0,2 und 0,3 – ähnlich wie der Zusammenhang zwischen Universitätsabschlussnote und Berufserfolg. Auch dieser Wert ist vergleichsweise hoch.

An dieser Stelle könnten viele weitere Kritikpunkte genannt werden, zum Beispiel:
- Die einer Note zugrundeliegenden Kriterien sind meist nicht in vollem Umfang bekannt.
- Eine Note reduziert die eigentliche Leistung auf eine Zahl und liefert keine differenzierende Möglichkeit zur Beurteilung der Leistungen.
- Noten sind zur Beurteilung bestimmter Sachverhalte wie Kreativität oder Sozialkompetenz nicht geeignet (Oelkers 2002).

Diese Argumente helfen jedoch bei der Professionalisierungsdebatte nicht weiter, da sie – streng genommen – für alle Formen der Bewertung gelten. Grundsätzlich hat sich das seit vielen Jahren gebräuchliche System der Ziffernnoten mit vier Abstufungen für „Bestanden“ und ein bis zwei Abstufungen für „nicht Bestanden“ bewährt. Oelkers (ebd.) nennt die wesentlichen Gründe dafür:
- Eine Skala von fünf oder sechs Noten erlaubt eine Beschreibung der Leistungsverteilung.

- [1] Unter einer Korrelation versteht man ein mathematisches Maß für einen Zusammenhang zweier Variablen. Dieses Maß ist eine Zahl zwischen –1 und 1. Eine Korrelation von 0 bedeutet dabei „kein Zusammenhang“, 1 ist ein perfekter positiver mathematischer Zusammenhang. Eine hohe Korrelation sagt aber überhaupt nichts über Kausalzusammenhänge aus, man muss die Kausalität immer theoriebasiert interpretieren.

- Die Beschreibung ist kurz und eindeutig.
- Das Notenschema ist ein öffentlicher Standard und wird nicht nur in der Schule verwendet.
- Das Schema lässt sich auf ökonomische Weise einsetzen und kommunizieren.
- Probleme der Ausdeutung sind gering.

Alternative Beurteilungsformen, die oftmals als Verbesserungsvorschlag genannt werden, weisen hingegen eine Reihe schwerwiegender Nachteile auf (Oelkers, ebd.):

- Textliche Leistungsbeschreibungen oder Wortgutachten verlangen einen erheblich größeren Aufwand, sind in stärkerer Weise interpretationsabhängig und haben Probleme vor allem bei der Formulierung negativer Beurteilungen. Sie verstecken die Urteile oft hinter differenzierter Freundlichkeit.
- Standardisierte Beurteilungsbögen verwenden zumeist weiche und vage Kriterien, zielen auf ganzheitliche Beurteilungen und müssen wiederum aufwändig interpretiert werden.
- Diskursive Verfahren, etwa Gespräche mit Eltern und Schüler/-innen, konfrontieren die Notengeber, also die Lehrkräfte, mit Akzeptanzproblemen, die nicht selten Machtproben darstellen. Alles, was begründet werden muss, verlangt erheblichen Aufwand und führt nicht immer zu einem glücklichen Ausgang.

Der Adressat der Kritik

Im Ergebnis ist jede Beurteilung ein Vergleich der individuellen Leistung mit einem Kriterium; dies gilt sowohl für Ziffernnoten als auch für mündliche Beurteilungen. Genau das ist es auch, was durch Noten ausgedrückt werden soll: Jede Note ist eine Einschätzung, ob und wie weit die gezeigte Leistung die erwarteten Anforderungen erfüllt oder nicht. Genau hierin liegt die wesentliche Funktion der Ziffernnoten. Die Kritik an ihnen ist damit streng genommen unberechtigt, da sie sich auf den gesamten diagnostischen Vorgang bezieht und dabei lediglich einen – zugegebenermaßen sehr prominenten – Teilbereich der Ergebnisrückmeldung benennt.

Dabei ist der Kerngedanke der Kritik, wie er exemplarisch in den Episoden 1 und 2 dieses Kapitels dargestellt wird, durchaus berechtigt: Wenn der Beurteiler kein professionelles Bewertungshandeln zeigt oder dieses nicht hinreichend transparent macht, vernachlässigt er die pädagogischen Potenziale der Leistungsbewertung, was zu negativen Konsequenzen für alle Beteiligten führt. Der Adressat der Kritik ist jedoch nicht die Ziffernnote, sondern die Lehrkraft, die unrichtig mit dem Instrumentarium der Leistungsbeurteilung umgeht. Ziffernnoten hingegen, und das sieht Oelkers vollkommen richtig, stellen im Vergleich zu anderen Formen ein relatives Optimum dar.

Daher ist auch, wie oben gezeigt wurde, das Ersetzen von Noten durch Verbalbeurteilungen meist keine Lösung (Episode 2). Außerdem würde dadurch der außerschulische Nutzen der Beurteilung stark eingeschränkt. Auch der Ehrgeiz der Eltern (Episode 1) hat nichts mit den Leistungen des Kindes zu tun; würden Sie nur deswegen bessere Noten vergeben, begingen Sie einen Wissen-um-die-Folgen-Fehler und würden evtl. dazu neigen, Kinder ehrgeiziger Eltern besser zu benoten als Kinder, deren Eltern keine weiteren Ambitionen haben. Das wäre jedoch von Grund auf falsch.

Was die inhaltliche Kritik am Punkteschema in Episode 1 angeht, muss man an dieser Stelle vorgreifen, dass der Gesetzgeber der Lehrkraft ein hohes Maß an professioneller Autonomie bei der Leistungsbewertung zuschreibt, solange die Zuteilung der Einzelleistungen zu Punkten bzw. zu Noten transparent und nachvollziehbar ist. Punkteschlüssel müssen also inhaltlich begründbar sein, müssen aber nicht verteidigt werden. Solange Sie also professionell und mit der nötigen Expertise gehandelt haben, als Sie ihre Punkteverteilung festgelegt haben, stellt sie kein Problem dar.

Episode 3 dieses Kapitels bezieht sich auf das in den Grundschulen vieler Länder verwendete Wortgutachten zur Beurteilung der Jahresleistung eines Lernenden. Hier lassen sich die geschilderten Nachteile und Probleme ausformulierter Beurteilungen gut illustrieren: Es gibt große Probleme bei der Deutung von Verbalbeurteilungen. Während jede Lehrkraft weiß, dass ein Schulzeugnis eben kein Arbeitszeugnis ist, ist den Eltern die Realität der Arbeitszeugnisse wegen der biographischen Nähe

zumeist deutlich stärker präsent und determiniert daher die Interpretation der Inhalte deutlich. Grundschulzeugnisse sollte man aber nicht wie Arbeitszeugnisse lesen; letztere müssen positiv formuliert sein, Grundschullehrkräfte werden sich dagegen einfach bemühen, taktvoll zu formulieren. Das bedingt aber nicht die Verwendung verschlüsselter Formulierungen, wie sie in Arbeitszeugnissen üblich sind. Das wäre bei Schulzeugnissen auch nicht sinnvoll, denn schließlich soll das Zeugnis ja über Lernprozesse und Lernerfolge (und evtl. auch Misserfolge) informieren. Dennoch machen die unterschiedlichen Interpretationen oft ein Gespräch über die Beurteilung nötig, da die Beteiligten über die Intention der Formulierung unterschiedlicher Meinung sind. Vielleicht meint eine Aussage wie „die Schülerin bemühte sich“ auch einfach: „Ihre Tochter strengt sich an.“ Bei Eltern sind derartige Missverständnisse noch auszuräumen, aber würde man wirklich Abschlusszeugnisse mit Verbalbeurteilungen in Betracht ziehen, wäre deren Verständlichkeit beispielsweise für einen potentiellen Arbeitgeber ein echtes Problem.

Die Vergabe von Noten ermöglicht die Vergabe von Ausbildungs- und Arbeitsplätzen oder Zugangsberechtigungen nach individueller Leistung. Sie bieten dabei ein hohes und gesellschaftlich anerkanntes Potenzial, einzelne Leistungen einzuordnen, zu vergleichen und zu vermitteln. Der gesellschaftliche Auftrag an die Schule geht jedoch weit über die hier kritisierte Selektionsfunktion hinaus (vgl. Fend 1980), sie erfüllt im Hinblick auf die Reproduktion der Gesellschaft und die Qualifikation des Einzelnen sehr wichtige Aufgaben, die nicht in der Rückmelde- und darauf aufbauenden Selektionsfunktion einer Ziffernnote abgebildet werden dürfen. Die Rolle der schulischen Leistungsbeurteilung für die Qualifikation und Sozialisation des individuellen Lernenden wird im nächsten Kapitel geklärt.

2.3. Bilanzierung des Gelernten

In diesem Kapitel werden folgende Lernimpulse gesetzt:

- Noten können im schulischen Kontext verschiedene Funktionen erfüllen, von denen einige kritisch zu sehen sind.
- Noten werden häufig kritisiert, doch ein Großteil der Vorwürfe geht entweder auf Fehler der bewertenden Lehrkraft zurück oder betrifft die Tatsache schulischer Bewertung generell.
- Neben Noten existieren noch weitere Formen schulischer Bewertung, etwa Verbalbeurteilungen, die jedoch zumeist mit deutlich höherem Aufwand verknüpft und ebenfalls nicht problemfrei sind.
- Dennoch gibt es auch gerechtfertigte Kritik, insbesondere an der Vergabepraxis von Noten. Sie sind nur dann ein gerechtes, vergleichbares Instrument zum Zweck der Information von Lehrkräften, Schüler/-innen und Gesellschaft, wenn subjektive Fehlerquellen bei der Benotung minimiert werden.

3. Pädagogische Aufgaben der Leistungsbeurteilung

Dieses Kapitel behandelt den Zusammenhang zwischen Leistungsbeurteilung und schulischer Förderung. Ausgehend von Überlegungen zu dem im letzten Kapitel erwähnten gesellschaftlichen Auftrag der Leistungsbeurteilung werden diagnostische Anforderungen abgeleitet, die für die individuelle Förderung von Schülerinnen und Schülern notwendig sind. Die Grundlage dieser Ausführungen bildet ein Leistungsbegriff, der sich sehr stark an pädagogischen Bildungsdefinitionen orientiert. Die daraus resultierenden Herausforderungen für das pädagogische Handeln werden mit den folgenden Episoden angerissen.

Episode 1: Fördern statt beurteilen

Grübeln ist nicht gut, das wissen Sie. Trotzdem gibt es einige Aspekte Ihres Unterrichts, die Sie auch nach der Schule nicht mehr loslassen. Vor allem die Aussage des Vaters eines ihrer Schüler macht sich in Ihrem Kopf breit. Das Kind tut sich schwer, das ist bekannt. Die Lösung des Vaters ist einfach: „Ich wünsche, dass Sie mein Kind ordentlich fördern. Es ist nicht hilfreich, wenn Sie es zensieren, denn damit werden die eigentlichen Qualitäten meines Kindes nur unzureichend dargestellt. Lassen Sie die Beurteilung weg, dann können Sie viel besser auf mein Kind eingehen. Ich weiß, dass das geht. Sie müssen es nur wollen." Sie sind im Zweifel. Unbewusst haben Sie den gleichen Gedanken schon mehrmals gehabt: Schadet die Leistungsbeurteilung der Förderung mehr, als dass sie nutzt?

Episode 2: Fördergespräch im Unterricht

Während einer Einzelarbeit im Fach Mathematik stellen Sie folgende Aufgabe: Zwölf Arbeiter haben bei neunstündiger Arbeitszeit in 7 Tagen 390 m² Betonschalung hergestellt.

Wie viele Arbeiter sind bei gleicher Leistung einzusetzen, wenn in insgesamt 21 Tagen 2 340 m^2 Betonschalung hergestellt werden müssen und die tägliche Arbeitszeit statt 9 nur noch 8 Stunden beträgt?

Sie bemerken, dass Max seine Aufgabe nicht löst, sondern Zeichnungen auf das Angabenblatt macht. Sie setzen sich neben Max und fragen nach:

Lehrkraft: Max, was ist los? Max: Ich hab keine Lust.

L: Aber du musst die Aufgabe machen. Verstehst du sie denn? M: Ja.

L: Warum löst du sie dann nicht? Du versuchst es ja gar nicht? M: Ich mag einfach nicht. Das ist doch blöd.

L: Soll ich dir helfen? M: Wenn Sie meinen …

L: Also schau her. Die Aufgabe ist ein dreifach verschachtelter Dreisatz: Proportional, antiproportional und nochmal antiproportional, wie wir es im Unterricht gemacht haben.

M: Hä?

L: Erstens: Je größer die einzuschalende Fläche ist, desto mehr Einschaler werden benötigt. Das ist proportional. Zweitens: Je mehr Zeit zur Verfügung steht, desto weniger Einschaler braucht man. Das ist antiproportional. Drittens. Je geringer die Arbeitszeit pro Tag ist, desto mehr Einschaler braucht man. Das ist auch antiproportional.

Jetzt brauchst Du nur noch die drei notwendigen Dreisätze ableiten und du kommst selbst auf die Lösung.

M: Und wie geht das?

L: Zuerst wird über die Fläche, dann über die Tage und dann über die Zeit geschlossen. Probier's mal!

M: Okay …

Max probiert es – und beginnt nach einer Minute wieder mit dem Malen. Was ist schiefgelaufen? Sie haben es ihm doch mathematisch korrekt erklärt …[2]

[2] Quelle der Aufgabe und der Lösungsansätze: Olaf Hinrichsen, http://www.brinkmann-du.de/mathe/fos/wieder02_01.htm

Episode 3: Was tun, wenn gar nichts mehr geht?
Maria hat keine Lust. Egal was Sie fragen, egal welche Angebote Sie machen, sie nimmt keine Aktivierung an und beteiligt sich nicht am Unterricht. Was können Sie tun?

3.1. Leistungsgesellschaft und Bildung

Um den Zusammenhang zwischen Leistungsbeurteilung und Förderung zu klären, soll zunächst absichtlich ein sehr einseitiges Bild gezeichnet werden: Viele der im letzten Kapitel genannten und diskutierten Anforderungen an die Leistungsbeurteilung zielen stark auf die Selektion der Schüler/-innen. Von mehreren Seiten aus werden (einseitig) die Kriterien der Leistungsgesellschaft auf die Schule angelegt und ihre daraus resultierende Aufgabe auf die schulische Leistungsbeurteilung fokussiert. Bildungstheoretisch berührt diese Diskussion die Frage nach Anerkennung und Implikationen der Leistungsgesellschaft. Deren Definition erscheint zunächst einfach:

„Unter Leistungsgesellschaft wird ein Sozialsystem verstanden, in dem Rollen und ihre Belohnungen in einem offenen Wettbewerb nach persönlicher Befähigung zugewiesen werden und Leistung als zentraler Wert gilt, der durch Sozialisation vermittelt und als Handlungsmotiv verinnerlicht wird. Geschlossene Sozialsysteme wie ständische und Kastengesellschaften sind die wichtigsten Antonyme. Sozialistische wie kapitalistische Industriegesellschaften verstehen sich als Leistungsgesellschaften; sie unterscheiden sich hinsichtlich der Bewertungsinstanzen (politisch-administrative Zentrale vs. Markt) und des Ausmaßes der Belohnungen im Schichtungssystem" (Endruweit/Trommsdorff, 1989).

Als Leistungsgesellschaften werden also – im Unterschied zu Ständegesellschaften – soziale Gruppierungen bezeichnet, in welchen der Erwerb von Funktionen und Ämtern, der soziale Status und materielles Einkommen auf der Grundlage erbrachter und bewertbarer persönlicher Leistung bestimmt werden. Daraus resultierende soziale Ungleichheiten werden unter Berufung auf die Leistung als strukturbestimmenden Verteilungsmaßstab allgemein akzeptiert. Die Leistungsgesellschaft zeichnet sich insbesondere durch flexible Übergänge zwischen den Schichten und die Möglichkeit steiler Karrieren aus (Jürgens 2005, S. 14).

Die Aufgabe, Schülerinnen und Schüler auf die Anforderungen der Leistungsgesellschaft vorzubereiten, wird in der gesellschaftlichen Diskussion zum großen Teil von der Schule gefordert. So schreibt beispielsweise die Bundesvereinigung deutscher Arbeitgeber in ihrem Positionspapier (1998):
„Die Schule hat die Aufgabe, durch Erziehung und Bildung junge Menschen zur Bewältigung der gesellschaftlichen Herausforderungen zu befähigen. Indem sie sowohl ein breites Grundlagenwissen als auch Schlüsselqualifikationen vermittelt, ermöglicht sie eine qualifizierte Berufsausbildung und ein qualifiziertes Hochschulstudium und legt die Grundlagen für eine lebenslange Weiterbildung. Sie trägt dabei mehr als bisher den Erfordernissen der Dienstleistungs-, Informations- und Wissensgesellschaft Rechnung. […] Schule als Teil des Gesamtsystems Gesellschaft muss sich den Anforderungen aus Wirtschaft und Gesellschaft öffnen, auf die Veränderungen des Umfelds reagieren und dieses Umfeld durch eigenes Tun mitgestalten."

Schlömerkemper (2002) nimmt diesen Gedanken auf und definiert den Lehrberuf mit Hilfe einer Kombination aus gesellschaftlichen Aufträgen. Einer dieser Aufträge ist es, Kinder und Jugendliche in ihrer Bildung zu unterstützen und es einem jeden, wie es auch in den Länderverfassungen steht, zu ermöglichen, „eine seinen erkennbaren Fähigkeiten und seiner inneren Berufung entsprechende Ausbildung zu erhalten" (Verfassung des Freistaates Bayern Art. 128 (1)).

Leistungsanforderungen in der Schule bilden damit, wenn man diesen Gedankengängen nur oberflächlich folgt, in erster Linie das gesellschaftliche Grundbedürfnis nach der Implementierung des Einzelnen in das wirtschaftliche und soziale System ab und eröffnen Möglichkeiten zur eigenen Entwicklung und Allokation. Im Hinblick auf die Bildungsdiskussion und auf das grundsätzliche Verständnis vom Lernen ist diese funktional ausgerichtete Begründung von Leistung und Leistungsbeurteilung jedoch viel zu kurz gegriffen und muss erweitert werden: Sie muss sich entfernen von einer reinen Fokussierung auf den

Output als messbares Ergebnis von Lernprozessen, welches für außerschulische Allokationszwecke wichtig ist und das wirtschaftliche Prinzip der Einstufung nach Leistung in die Schule hineinträgt. Stattdessen muss man Leistung und Leistungsbeurteilung unabhängig von den Leistungsanforderungen der Gesellschaft auch für die Entfaltung der eigenen Individualität, die Entstehung von Mündigkeit und Selbstständigkeit, die Erschließung der Möglichkeiten für den Einzelnen und das produktive Mitgestalten unserer Gesellschaft nutzbar machen. In diesem Verständnis ist der Leistungsbegriff vom Bildungsbegriff kaum noch zu unterscheiden (Schlömerkemper, ebd.).

Dieses Verständnis von Leistung bildet die Grundlage dieses Bandes. Es geht damit bei der Leistungsbeurteilung nicht nur um die Outputmessung, sondern auch um die Nutzbarmachung diagnostischer Modelle als Mittel zur Förderung der Lernprozesse von Schüler/-innen. Dieser Meinung sind auch die Kultusministerien der Bundesrepublik Deutschland. In den Standards für die Lehrerbildung im Bereich Bildungswissenschaften wird der Kompetenzbereich *Diagnostizieren und Leistung messen* definiert als:

„Lehrerinnen und Lehrer erfassen Leistungen von Schülerinnen und Schülern auf der Grundlage transparenter Beurteilungsmaßstäbe."

Das bedeutet unter anderem: „Die Absolventinnen und Absolventen kennen die Prinzipien der Rückmeldung von Leistungsbeurteilung und begründen Bewertungen und Beurteilungen adressatengerecht und zeigen Perspektiven für das weitere Lernen auf" (KMK 2004).

An dieser Stelle sei kurz auf die Fragestellung der Episode 1 verwiesen. In ihr wird nach der Bedeutung der Leistungsbeurteilung für die Förderung gefragt, wobei nicht nur ihre Bedeutung in Frage gestellt wird, sondern Leistungsbeurteilung und Förderung sogar als Gegensatz gesehen werden. Solche Überlegungen sind jedoch nicht korrekt, da die Leistungsbeurteilung die wichtigste Voraussetzung für eine erfolgreiche Förderung ist. Die folgende Begründung dieser Feststellung ebnet zugleich den Weg zu einer adäquaten Herangehensweise an die entsprechenden Handlungen.

Einen ersten Ansatzpunkt liefert das in Kapitel 1 vorgestellte Handlungsregulationsmodell von Frese und Zapf: Inwiefern der Lernende seine Lernhandlung ändert, entscheidet er letztendlich selbst. Die Wahrscheinlichkeit hierfür kann jedoch durch die Qualität des Feedbacks, also durch die Qualität des Förderimpulses, gesteigert werden. Wenn Leistungsbeurteilung zu einem förderorientierten Feedback führt und ihre Ergebnisse in handlungsorientierte Anleitungen überführt werden, dient sie unmittelbar der Förderung der beurteilten Lernenden.

Diese Erkenntnis ist einfach zu verstehen, jedoch anspruchsvoll in der Umsetzung, weshalb an dieser Stelle genauer auf diesen Prozess eingegangen werden muss.

3.2. Leistungsbeurteilung als Basis der Förderung

Die folgenden Ausführungen erweitern, differenzieren und präzisieren den bereits eingeführten pädagogischen Handlungsraum des diagnostischen Kreislaufs aus Kapitel 1. Hierzu wird zunächst das Ziel dieses Kreislaufs, die Veränderung von Lernprozessen zur Erweiterung von Wissen und Kompetenzen, nochmals aufgenommen und anhand von Vygotskys Theorie zur Zone der proximalen Entwicklung präzisiert.

Vygotskys Zone der proximalen Entwicklung

Obwohl die Theorien von Lev Vygotsky (1978) ihren Ursprung in den 20er Jahren des letzten Jahrhunderts haben und formal gesehen vor allem auf den – für die moderne Pädagogik eher unergiebigen – Arbeiten von Iwan Pawlow aufbauen, steht Vygotskys Lernbegriff für eine Theorie, die 50 Jahre später unter dem Namen „Konstruktivismus" bekannt wurde. Im Kern erkennt Vygotsky nämlich an, dass der Lernende nicht durch Informationsdarstellung eines erfahrenen Menschen, also zum Beispiel durch eine wohlformulierte Unterweisung, sondern durch eigenes Handeln in typischen Problemfeldern neue Kompetenzen erwirbt. Er setzt also die Handlungen des Lernenden in den Mittelpunkt seiner Betrachtungen und definiert in

diesem Kontext drei Handlungsbereiche, die in der folgenden Abbildung illustriert werden.

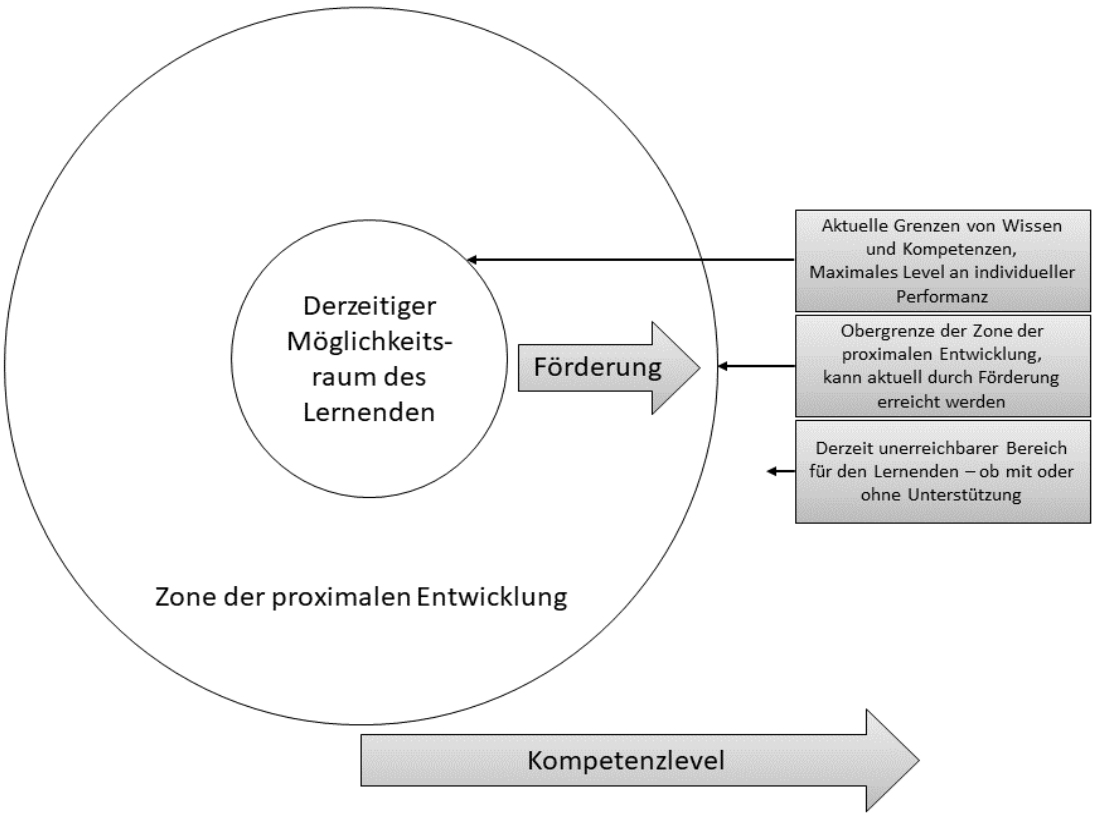

Abb. 15: Illustration der Zone der proximalen Entwicklung nach Vygotsky

Handlungsbereich 1: Derzeitiger Möglichkeitsraum. Dieser Bereich ist definiert als der aktuelle Möglichkeits- oder Handlungsraum des Lernenden, also der Bereich, in dem der Lernende Problemstellungen und Herausforderungen unabhängig von unterstützenden Hilfestellungen bearbeiten und lösen kann. Es ist ihm jedoch nicht immer möglich, diesen Bereich alleine durch selbstständiges Arbeiten und Entdecken zu erweitern, da er an Verständnisgrenzen stößt, die ohne weitere Informationen oder Hilfestellungen nur schwer zu durchbrechen sind. Mitunter ist es didaktisch auch nicht sinnvoll oder sogar gefährlich, den Überblick und das angemessene Handeln in einem erweiterten Raum selbst zu entdecken (z. B. im Sport: richtig fallen lernen).

Handlungsbereich 2: Die Zone der proximalen Entwicklung. Vygotsky beschreibt daher die Zone der proximalen Entwicklung als einen Möglichkeitsraum, innerhalb dessen sich ein Lernender weiterentwickeln kann. Er kann darin Probleme zwar nicht mehr – wie in Zone eins – selbstständig lösen, ist dazu im Gegensatz zu Zone drei jedoch in der Lage, wenn er in seinen Bemühungen unterstützt wird. Daher ist es diese Zone, die der Unterricht besonders adressieren sollte. Letztendlich sollte diese Zone durch Übung nach und nach zu einem erweiterten Möglichkeitsraum des Lernenden werden.

Hierzu führt der Lehrende den Lernenden zunächst durch die Zone der proximalen Entwicklung und nutzt dabei beispielsweise Stichworte, Hinweise, Modellbildung, Erklärungen, Leitfragen, Diskussionen, Mitwirkung, Ermutigung, Aufmerksamkeitssteuerung etc. – er baut gewissermaßen Brücken zwischen den vorhandenen Fähigkeiten des Kindes und neuen Fertigkeiten. Vykotsky nimmt dabei an, dass diese Unterstützung umso erfolgreicher ist, je vertrauter der Lehrende selbst mit dieser Zone der proximalen Entwicklung ist (dies ist – nebenbei bemerkt – auch eine Begründung für den Erfolg von Tutorien und Lernpartnern/Lerngruppen). Das verlangt von dem Lehrenden im Ergebnis eine sprachliche Anpassung an den Wissensstand des Kindes, die behutsame Einführung neuer Handlungsmöglichkeiten und Zusammenhänge sowie die Kenntnis von erfolgreichen Lern- und Verständniswegen für diesen Lernenden.

Handlungsbereich 3: Die Zone außerhalb der proximalen Entwicklung. Dieser Bereich ist für den Lernenden mit seinem derzeitigen Wissensstand unerreichbar – sei es mit oder ohne Hilfestellung. Der Lernende hat keine Möglichkeit, diesen Bereich zu durchdringen, es sollte also bei der Förderung gar nicht versucht werden, diesen zu erreichen, bevor nicht die Zone der proximalen Entwicklung vom Lernenden selbstständig mit Handlungen ausgefüllt werden kann. Illustrieren kann man diese Zone beispielsweise mit Fremdwörtern: Wo eine erfahrene Lehrkraft Fremdwörter wie *Interaktion*, *Devianz* oder *Determinismus* gut versteht und mit eigenen Bedeutungen versieht, hat ein unerfahrener Lernender keinen Bezug dazu – und kann diesen auch nicht unmittelbar bilden, da ihm die dazu notwendigen Vorkenntnisse und Anschlüsse nicht zur Verfügung stehen.

Diese Erkenntnis ist die Basis für eine optimale Gestaltung der Förderung. Im Endeffekt handelt es sich dabei um nichts anderes als eine Hilfestellung für den Lernenden mit dem Ziel, dass dieser sich die Zone der proximalen Entwicklung als seinen nächstgrößeren Möglichkeitsraum erwirbt. Das eröffnet eine Reihe von Herausforderungen, die man nur bewerkstelligen kann, wenn man den Lehrberuf als Expertise definiert:

- Die professionell arbeitende Lehrkraft kennt die Zusammenhänge und Hierarchien ihres Faches, sie kann also, aufgrund ihrer Expertise, sehr genau schließen, welcher Handlungsraum als Zone der proximalen Entwicklung definiert werden kann.
- Diese Zone ist aber höchst individuell und wird zu einem großen Teil vom derzeitigen Wissens- und Entwicklungsstand des Lernenden bestimmt. Daher benötigt der Lehrende, neben seinem fachlichen und didaktischen Wissen, unbedingt die richtigen diagnostischen Kompetenzen, um den derzeitigen Möglichkeitsraum eines Lernenden möglichst korrekt zu diagnostizieren.
- Insbesondere ist es wichtig, die Zone der proximalen Entwicklung zu erkennen und zu nutzen. Denn wenn die Lernhilfen auf dem aktuellen Entwicklungsniveau liegen, lernt der Lerner nichts dazu, liegen sie außerhalb der Zone der nächsten Entwicklung, ist der Lerner überfordert und reagiert frustriert.

Dies illustriert übrigens einen der Kardinalfehler des Förderns: Oftmals sind die Fördermaßnahmen zu standardisiert und lehnen sich nicht an den eigentlichen Bedarf des Lernenden an. Den derzeitigen Möglichkeitsraum zu erkennen und die notwendigen Fördermaßnahmen der Zone der proximalen Entwicklung anzupassen, erfordert zunächst einmal hohe diagnostische Kompetenzen.

Das BIB-Modell der Diagnostik

Um die Zone der proximalen Entwicklung zu diagnostizieren, muss man sich zunächst klar darüber sein, wo die Grenzen des derzeitigen Möglichkeitsraums des Lernenden liegen. Dies zu erfahren ist nicht leicht. Ein alleiniger schriftlicher Wissenstest greift hier zu kurz, da dieser zu sehr auf das Ergebnis zielt und Gründe für Leistungslücken vor allem auf der fachlichen und der Anwendungsebene gefunden werden können. Weitere, auch persönliche Aspekte des Lernenden, kommen hier zu kurz. Damit bleibt als Erfahrungsgrundlage vor allem das Gespräch mit den Schüler/-innen. Doch auch hier ist es nicht einfach und nur mit der korrekten Fragetechnik möglich, Entwicklungsstände zu erfahren.

Auf sicherem Wege erreicht man die notwendigen Kenntnisse mit lästigem und langandauerndem Nachbohren. Das ist im schulischen Kontext jedoch meist keine gute Idee, endet doch jeder derartige diagnostische Prozess, ähnlich wie in der leichtathletischen Disziplin des Hochspringens, mit einer gefühlten Niederlage für den direkt Beteiligten: Wie erkenne ich, wie hoch ich springen kann? Indem ich dreimal reiße. Wie erkenne ich die Grenzen meines Möglichkeitsraumes? Indem ich so lange antworte, bis die Lehrkraft mehrmals rückmeldet, dass meine Antwort falsch ist. Dieses Vorgehen ist im Allgemeinen unbefriedigend.

Besser ist es, die Grenze des derzeitigen Möglichkeitsraumes durch die diagnostische Anwendung des bereits bekannten BIB-Modells (Kiel 2001) zu erkennen. Die Grundidee dieses Vorgehens liegt in der Annahme, dass echtes Verständnis nur gegeben ist, wenn ein spezifisches Thema in die Lebenswirklichkeit des Lernenden integriert ist (Merrill 2002). Das ist immer dann der Fall, wenn der Lernende zu spezifischen Theorien eigene Anwendungen nennen und die Nutzbarmachung des Themenbereiches für

eigene Problemstellungen beschreiben kann. Konkret kann dieses Vorgehen mit Hilfe der folgenden Grafik illustriert werden:

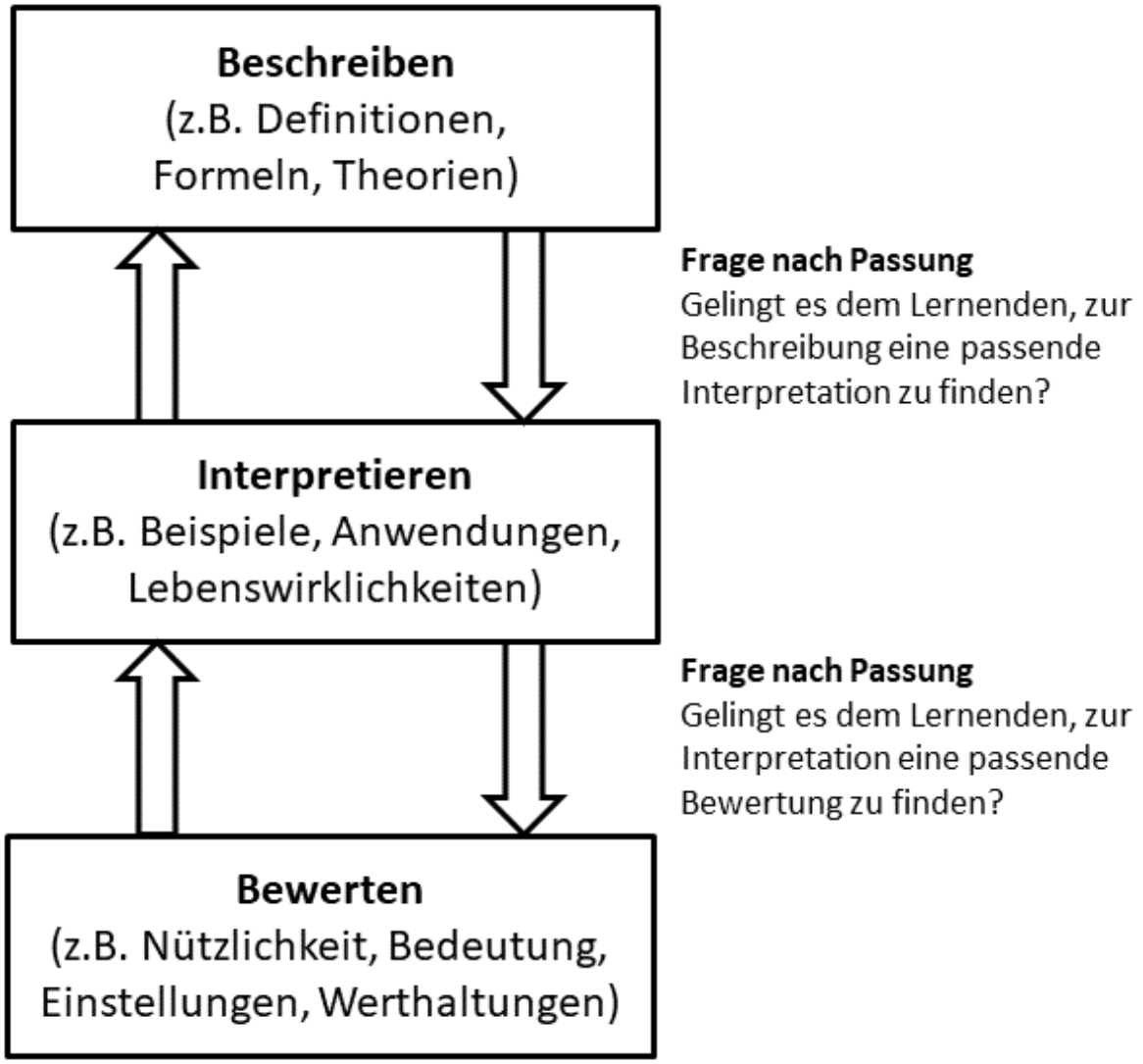

Abb. 16: Das BIB-Modell der Diagnostik

Bei der Befragung zielt man daher darauf ab, die Ebenen „Beschreibung", „Interpretation" und „Bewertung" aus der Sicht des Lernenden kennen zu lernen (vgl. Kapitel 1 und 7 dieses Buches).

Beschreiben. Der Terminus *Beschreiben* adressiert die typischen Verständnis- und Behaltensleistungen – und er ist häufig zu dominant in diagnostischen Prozessen. Alleine aus dem Wissen und dem Verständnis eines Lernenden lassen sich die Grenzen seines derzeitigen Handlungsraumes nicht ableiten, dazu sind die potenziellen Fragestellungen zu fachorientiert und nicht genügend auf die damit verbundenen Handlungen bezogen. Dennoch sind Fragestellungen, die auf das Beschreiben von Phänomenen zählen, wichtig. Beispielsweise könnte gefragt werden: Wie lässt sich ein Gegenstandsbereich definieren? Welche Theorien und Axiome liegen diesem Gegenstandsbereich zugrunde? Was sind die Zusammenhänge zwischen zwei Aspekten? Kannst Du das Ganze in eigenen Worten beschreiben?

Interpretieren. Hier geht es um die individuellen Deutungen eines Themas sowie um die mögliche Anwendung der Inhalte für vielfältige Aufgaben. Interpretationsfragen suchen demnach nach der Nutzbarmachung des Themenbereichs für typische Herausforderungen innerhalb und außerhalb der Schule sowie nach Beispielen und Begründungen, fragen also nach Transfer-, Analyse- und Syntheseleistungen. Auch diese Fragen werden im Unterricht gestellt, jedoch nicht immer mit der gebotenen Häufigkeit.

Beispiele hierfür sind: Warum sind die Dinge so, wie sie sind oder gemacht werden? Mit welchen Beispielen kann man das illustrieren? Welche Anwendungen sind möglich? Wie kann man das Themengebiet in der Praxis nutzen?

Bewerten. Beim Bewerten geht es um affektive Fragestellungen, die wichtig sind, um den Bezug des Lernenden zum Thema zu erfahren. Mit ihnen kann man prüfen, auf welchen Wegen die Bereitschaft, sich tiefergehend mit den Inhalten zu beschäftigen, angestoßen werden kann. Bewertungsfragen werden üblicherweise zu selten gestellt, dabei sind gerade sie es, die erste Ansätze zur Förderung eröffnen, da sie an der Werthaltung des Lernenden zum aktuellen Themenbereich andocken: Man kann die – für den Lernprozess äußerst wichtigen – Erwartungen und Einstellungen des Lernenden nur dann für die Förderung nutzen, wenn man sie kennt. Ansonsten bleiben viele Unterstützungsangebote für den Lernenden bedeutungsfrei und damit unpassend. Mögliche Fragekomplexe sind hier: Welche Bedeutung hat das Thema für den Lebensbereich? Ist das Thema für den Lernenden wichtig oder unwichtig? Kann sich der Lernende vorstellen, wozu das Thema wichtig und bedeutsam sein könnte – auch in Zukunft? Ist diese Anwendung wichtig für ihn?

Wie dieser diagnostische Prozess initiiert wird, ist an dieser Stelle wegen der kontextuellen Reichhaltigkeit des Schulstoffes nicht direkt definierbar. Das übliche Vorgehen, mit einer einfachen Inhaltsfrage zu beginnen, muss auch für diese Form der Leistungserhebung nicht abgelehnt werden. Und selbstverständlich hält sich der Lehrende auch bei solcherart Befragungen an die bekannten Regeln, stellt also beispielsweise klare und verständliche Fragen, springt bei der Befragung nicht und denkt daran, bei der Befragung nicht zu belehren. Wichtig ist jedoch vor allem, dass man die Antworten des Lernenden den drei Kategorien zuordnen kann, man also erkennt, ob sie/er beschreibt, interpretiert oder bewertet – und in welcher Qualität sie/er das macht.

Wenn eine Frage den derzeitigen Möglichkeitsraum des Lernenden anspricht, sind die Antworten in allen drei Aspekten (Beschreibung, Interpretation, Bewertung) sicher und präzise. Wann immer die Antworten des Lernenden unpräzise werden, ohne zusätzliche Hilfe keine Lösungen gefunden werden oder die Interpretation/Bewertung nicht zur Beschreibung passt, ist dies ein Hinweis darauf, dass die Inhalte aus der Zone der proximalen Entwicklung stammen können. Beispiele hierfür sind:

- Worthülsen, die erkennen lassen, dass der Lernende die Bedeutung nicht wirklich begriffen hat,
- Antworten, die erkennen lassen, dass sich der Lernende erst zum eigentlichen Ergebnis hin entwickeln muss,
- Beispiele, die nicht mit Klarheit und Umsetzbarkeit überzeugen,
- Aufzählungen, die zwar irgendwie zum Thema gehören, aber nicht auf den Punkt kommen,
- Versuche, von der Fragestellung abzulenken (alltagssprachlich: Herumeiern) aber auch
- auswendig gelernte Begrifflichkeiten ohne eigenes Verständnis. Hier ist es hilfreich, nach Interpretationen und Beispielen zu fragen.

Grundsätzlich kann man festhalten: Die Grenze zwischen dem momentanen Möglichkeitsraum und der Zone der proximalen Entwicklung liegt da, wo die Antworten von Klarheit in Vagheit übergehen. Genau hier liegt auch der Ansatzpunkt für die individuelle Förderung.

Der Vollständigkeit halber sei noch erwähnt, dass die obere Grenze der proximalen Entwicklung dort liegt, wo keine Antworten mehr gegeben werden können, da die Fragen Gegenstandsbereiche berühren, mit denen der Lernende nichts anzufangen weiß.

Die Expertise des Ausfragenden ist dabei essentiell für die erfolgreiche Diagnostik von Wissens- und Verstehenslücken. Dabei ist nicht nur die fachliche Expertise relevant, sondern auch die Kenntnis der möglichen Wissenslücken von Schülerinnen und Schülern – und die Kompetenz, richtig zu fragen. Diese Kompetenz erwirbt man durch das Verständnis und die Anwendung der genannten Faustregeln, durch viel Erfahrung und selbstständige Reflexionsprozesse.

Konsequenzen für die Förderung

Der beschriebene Weg, die Grenzen zwischen dem derzeitigen Möglichkeitsraum und der Zone der proximalen Entwicklung eines Lernenden zu erkennen, eröffnet für die unterrichtliche Praxis eine Reihe von Handlungsmöglichkeiten. Natürlich ist diese Operationalisierung für alle Beteiligten mit Anstrengungen verbunden. Dennoch, und das zeigen auch die folgenden Erkenntnisse der Lehr-Lern-Forschung, bilden die Grundgedanken dieses Kapitels einen gangbaren Weg, individuelle Förderung anzustoßen. Konkret führt das zu folgenden Forderungen:

- **Häufige formative Evaluation des Unterrichts.** Laut Hattie (2013) ist eine der wichtigsten Bedingungen für die Förderung des Lernprozesses die permanente formative Evaluation des Unterrichts. Das bedeutet: Es ist notwendig, fortwährend zu überprüfen, wie es um den Wissensstand, die Einstellungen und die Bereitschaft der Lernenden bestellt ist – im Idealfall möglichst individuell nach jeder erfolgreich durchlaufenen Lerneinheit.
- **Intensivierung der Interaktion mit den Lernenden.** Ein weiterer Hinweis auf die Notwendigkeit der pädagogischen Diagnostik kommt aus der Conceptual-Change-Forschung. Diese geht von der Annahme aus, dass Bereitschaft und Lernprozess zu einem großen Teil von den Einstellungen und Vorkenntnissen der Lernenden determiniert werden. Um Lernprozesse anzustoßen, müssen nach Sinatra und Mason (2008) die Barrieren, die einem Perspektivwechsel entgegenstehen, umfangreich bestimmt und die neuen Lernprozesse thematisiert werden. Außerdem sollten intensive Interaktionsprozesse und aktives Experimentieren stattfinden. Dies gilt umso mehr, je geringer die Bereitschaft des Lernenden ist, an den eigenen Konzepten etwas zu ändern.
- **Den Unterricht aus den Augen des Lernenden sehen.** Dieser Punkt ist für Hattie (ebd.) die wichtigste Konsequenz aus seinen Metastudien. Diagnose dient nicht in erster Linie dazu, die Leistung des Lernenden zu beurteilen, sondern um zu erkennen, wie der Lernende die Unterrichtsthemen und -gegenstände wahrnimmt und versteht, um darauf aufbauend Hilfen und Unterstützungen anzubieten.

Weitere interessante Hinweise auf die Gestaltung der Förderung gibt Band 1 dieser Reihe im Kapitel 5.3.

In Bezug auf Episode 1 sollten Sie jetzt wissen, dass eine fachgerechte Leistungsbeurteilung grundsätzlich auch der Förderung dient und nicht, wie vom Vater ausgedrückt, ihr zwangsweise entgegensteht. Inwiefern Sie in diesem Fall von Zensuren abrücken dürfen, erfahren Sie in Kapitel 4. In Kapitel 5 stoßen Sie auf Möglichkeiten, wie Sie ein schwaches Kind im Rahmen der Leistungsbeurteilung zusätzlich fördern und schlechten Zensuren ihren Schrecken nehmen können.

Ein Beispiel für eine ungenügende Hilfestellung in der Zone der proximalen Entwicklung liefert hingegen Episode 2. Der Lehrer hat korrekt erkannt, dass Max die Aufgabe vermutlich nicht verstanden hat. Er erkennt jedoch nicht, dass er eine Frage gestellt hat, die bereits außerhalb von Max' Zone der proximalen Entwicklung liegt – und er deshalb nichts damit anfangen kann. Eine positive Antwort auf die Frage *Verstehst du sie denn?* bedeutet noch nicht, dass Max die Aufgabenstellung oder gar den Lösungsalgorithmus tatsächlich verstanden hat. Der Lehrer wählt darüber hinaus sein Hilfsangebot, ohne das Vorverständnis von Max zu prüfen, wodurch er erneut seine Hilfestellung nicht innerhalb, sondern außerhalb der Zone der proximalen Entwicklung anbietet. So kann Max nichts damit anfangen. Ähnliche Probleme sind übrigens auch in Fächern verbreitet, die nicht primär kognitive Leistungen abprüfen. In Sport etwa wird es einem Kind nur wenig bringen, zu wissen, dass es zu langsam ist oder den Ball nicht trifft. Entscheidend ist, warum es so ist und was es dagegen machen kann. Das gleiche gilt etwa in Kunst, wenn es einem Prüfling nicht gelingt, eine bestimmte Zeichenaufgabe zu erfüllen.

Ein besserer Einstand wäre – nicht nur in diesem Beispiel – die Aufforderung an Max, die Problemstellung mit eigenen Worten zu beschreiben oder den Zusammenhang zwischen dem Dreisatz und der Aufgabenstellung zu formulieren. Bei diesen Versuchen wären die Grenzen seines Verstehens deutlicher zu erkennen gewesen.

Bevor nun der Rahmen der gesetzlichen Vorgaben, die das diagnostische Handeln einer Lehrkraft bestimmen, genauer definiert wird, sei an dieser Stelle noch ein Hinweis gestattet, der sich auf Episode 3 bezieht: Nicht immer liegen die Gründe für nicht vorhandene Antworten in der mangelnden Kompetenz der Lernenden. Manchmal gibt es andere Gründe. Auch diese müssen Sie als verantwortungsvoller Pädagoge ermitteln und daraus die gebotenen Konsequenzen zum Wohl des Lernenden ziehen – auch wenn diese mitunter nicht einmal etwas mit dem Unterricht zu tun haben. Vielleicht hat Maria einfach nur Liebeskummer.

3.3. Bilanzierung des Gelernten

In diesem Kapitel werden folgende Lernimpulse gesetzt:

- Es greift zu kurz, Leistungsbeurteilung in der Schule nur als Erfordernis und Abbild der Leistungsgesellschaft zu verstehen. Sie kann vielmehr dem Lernenden auch bei der Entwicklung von Individualität, Mündigkeit und Selbstständigkeit helfen.
- Leistungsbeurteilung ist eine wesentliche Grundlage schulischer Förderung. So ist sie etwa in Vygotskys Zonenmodell eine Grundbedingung, ohne die die Lehrkraft nicht wissen könnte, wo die Zone der proximalen Entwicklung und damit der ideale Förderbereich des einzelnen Lernenden liegt.
- Die Zone der proximalen Entwicklung kann man am besten im Gespräch mit Schüler/-innen verorten. Eine Möglichkeit besteht darin, nach dem BIB-Prinzip zu fragen und so zu überprüfen, inwiefern Schüler/-innen gelerntes Wissen auch interpretieren und für sich selbst bedeutsam machen können.
- Um Diagnostik zur Förderung von Lernenden einsetzen zu können, muss man den Unterricht möglichst fortlaufend formativ evaluieren, die Interaktion mit den Schüler/-innen intensivieren und versuchen, den eigenen Unterricht aus den Augen der Lernenden zu sehen. So ist man in der Lage, aus den Antworten der Lernenden auf ihren Förderbedarf zu schließen und die Lehr-Lern-Situationen daran anzupassen.

4. Gesetzliche Bestimmungen der Leistungsbeurteilung

Das nächste Kapitel geht einen weiteren Schritt hin zum professionellen Umgang mit den Anforderungen der Leistungsbeurteilung, indem es den grundlegendsten und nicht veränderbaren Rahmen definiert, innerhalb dessen das Handeln als Diagnostiker stattfindet: die Gesetze und Verordnungen. Da diese nicht in allen Ländern gleich sind, prägen sie auch ein gutes Stück weit die Kultur der Leistungsbeurteilung eines Landes. Viele der bislang angesprochenen Herausforderungen, etwa die Notwendigkeit einer sachbezogenen Bezugsnorm, haben beispielsweise im Schweizer Schulsystem eine deutlich längere Tradition und damit eine gründlichere Implementation in die Bewertungskultur an Schulen (BKZ 2013). Grundlage dieses Buches bildet das derzeit gültige Schulrecht in den Ländern der Bundesrepublik Deutschland. An dieser Stelle muss angemerkt werden, dass die hier dargestellten Rechtsnormen ständiger Veränderung unterworfen sind und damit eine Allgemeingültigkeit der Darstellungen bzw. eine ständige Korrektheit nicht gewährleistet werden kann.

Das Kapitel behandelt deutschlandweit geltende Regeln und Ausschnitte aus der Ländergesetzgebung sowie aus Verordnungen. Darüber hinaus nennt es Beispiele für Fälle, zu denen diese Vorgaben schweigen, bei denen die Lösung also innerhalb der Schule oder in Zusammenarbeit mit den Eltern gesucht werden muss. Es geht auch auf Fragen zu Widersprüchen und Beschwerden gegen Notenentscheidungen ein.
Im Folgenden werden häufige Problemstellungen zu den gesetzlichen Bestimmungen anhand von drei Episoden illustriert:

Episode 1: Eine ungerechte Frage
Der Schüler einer Freundin erzählt Ihnen folgende Geschichte: „In einem mündlichen Ausfragen in Physik ließ mich mein Lehrer die Frage beantworten, ob man, wenn man vor einem Spiegel zurücktritt, einen größeren oder kleineren Ausschnitt seines Körpers (in der Senkrechten) sieht. Mit Hilfe der entsprechenden Formeln konnte ich nachweisen, dass die sichtbaren Bereiche gleich bleiben. Seine Nachfrage, warum man dann im Allgemeinen gerne für einen letzten prüfenden Blick in den Spiegel einen Schritt zurücktritt, konnte ich nicht beantworten und bekam deshalb nur eine 2. Ich finde diese Frage ungerecht, denn sie hat nichts mit dem gelernten Stoff zu tun, im Unterricht haben wir das auch nicht durchgenommen und die eigentliche Frage war richtig beantwortet." Was raten Sie ihm?

Episode 2: Diskussionen bei der Klassenarbeit
Sie geben eine schriftliche Mathematikarbeit zurück. Der Schnitt ist sehr schlecht ausgefallen. Ein Drittel der Klasse hat die Note 5 und der Durchschnitt liegt auch nur bei 4,1. Es entspannt sich eine rege Diskussion. Ein Schüler äußert die Meinung, dass die Schulaufgabe damit nicht gültig ist. Wenn der Schnitt unter 4,0 liegt, muss sie zurückgenommen werden. Das habe er irgendwo gelesen. Eine andere Schülerin beschwert sich: Sie hat von 44 Punkten 23 Punkte erreicht. Ihres Erachtens wäre das eine 3. Außerdem beklagen sich die Schüler, dass Ihnen nicht die volle Zeit von 45 Minuten zur Verfügung stand, sondern mindestens 10 Minuten weniger. Drei Viertel der Klasse habe die Aufgaben wohl aus diesem Grund nicht geschafft.

Danach gehen Sie die einzelnen Aufgaben mit den Schüler/-innen durch. Bei Aufgabe 3 schreit plötzlich Elke durch die Klasse: „Frau Lehrerin, bei meiner Nachbarin haben sie einen Fehler übersehen. Den gleichen Fehler habe ich gemacht, aber mir haben sie ihn angestrichen. Deswegen habe ich eine 5. Und die hat jetzt eine 4. Das ist ungerecht. Ich will auch eine 4. Sonst fall ich durch!"
Insgesamt eine sehr unglückliche Situation. Wie ist die Rechtslage hierzu?

Episode 3: Die Dienstaufsichtsbeschwerde
Ihre Rektorin bittet Sie zu einem Gespräch. Es liegt eine Dienstaufsichtsbeschwerde gegen Sie vor. Ein Schüler möchte Ihre Korrektur einer Probearbeit im Fach Deutsch anfechten. Im Kern formuliert er seine Vorwürfe wie folgt: „Ich glaube, sie hat die letzte Deutschschulaufgabe so bewertet, dass sie nur das als gut beurteilt hat, was ihrer Meinung entspricht. Andere Meinungen lässt sie nicht zu." Wie reagieren Sie am besten?

4.1. Deutschlandweite Regelungen

Das so genannte Schulrecht ist eine Sammelbezeichnung aller Rechtsnormen und Verordnungen, die die Schule bzw. die mit dieser Institution verbundenen Personen betreffen. Es regelt insbesondere die mit dem Schulbetrieb zusammenhängenden Rechte und Pflichten der Beteiligten (z. B. Lehrer/-innen, Schüler/-innen, Eltern, Schulleitung oder Schulaufsicht).

Der oft gehörte Satz „Bildung ist Ländersache" wird indirekt durch den Artikel 30 des Grundgesetzes der Bundesrepublik Deutschland bestimmt:

„Die Ausübung der staatlichen Befugnisse und die Erfüllung der staatlichen Aufgaben ist Sache der Länder, soweit dieses Grundgesetz keine andere Regelung trifft oder zulässt." Da das Grundgesetz keine Regelungen definiert, die über Artikel 7[3] *hinausgehen (man sagt dazu: Das Grundgesetz schweigt in Fragen des primären und sekundären Bildungssektors), ist Schulrecht in Deutschland Länderangelegenheit. Die einzelnen Gesetze und Verordnungen sind jedoch länderübergreifend sehr ähnlich und werden im folgenden Kapitel durch die Bestimmungen des Bayerischen Schulrechts illustriert.*

Zur Einheitlichkeit des deutschen Schulrechts tragen neben gemeinsamen Traditionen vor allem die Absprachen und förmlichen Vereinbarungen der durch Staatsvertrag zwischen den Ländern eingesetzten Kultusministerkonferenz sowie die gegenseitige Anerkennung von Schulabschlüssen und Lehrbefähigungen bei.

4.2. Regeln in den Landesgesetzen

Gesetze und Verordnungen sind in der Regel hierarchisch aufgebaut. Die Länderverfassungen nehmen den Auftrag des Grundgesetzes als nächste gesetzliche Hierarchieebene hinsichtlich der staatlichen Aufsicht von Bildung, Erziehung und Unterricht auf und definieren im Wesentlichen zunächst

- die generellen Bildungs- und Erziehungsziele (wie z. B. Mündigkeit, Selbstständigkeit, Sozialisation, Integration und Emanzipation),
- den Auftrag zur ganzheitlichen Bildung (nicht nur Kopf, sondern auch Herz und Hand),
- übergeordnete Lernziele (z. B. Erziehung zum Geiste der Demokratie, Verantwortungsgefühl, Hilfsbereitschaft oder weitere Sozialkompetenzen),
- sowie die Erziehung zur Mitwirkung in der Sozialgemeinschaft oder der Familie.

Auf der nächsten Hierarchiestufe stehen Gesetze, die generelle Regelungen über das landesweite Erziehungs- und Schulwesen treffen. In ihnen werden die grundsätzlichen Regelungen für die institutionelle Bildung im Rahmen des staatlichen Schulwesens definiert. Dies betrifft auch die schulische Leistungsbeurteilung. Dabei beschreiben die Gesetze in einem ersten normativen Schritt eine generelle Verpflichtung zur Leistungskontrolle an Schulen; so beginnt beispielsweise der Artikel 52 des Bayerischen Gesetzes über das Erziehungs- und Unterrichtswesen (BayEUG) mit dem Passus:

„Zum Nachweis des Leistungsstands erbringen die Schüler in angemessenen Zeitabständen entsprechend der Art des Fachs schriftliche, mündliche und praktische Leistungen."

Das bedeutet konkret: Unterricht kann und darf nicht ohne Leistungsbeurteilung umgesetzt werden, es besteht die Pflicht des Lehrenden, die Kompetenzen der Lernenden zu evaluieren und das Ergebnis rückzumelden. Der Gesetzgeber erkennt damit die in den vorherigen Kapiteln dargestellte Notwendigkeit der Leistungsbeurteilung zur erfolgreichen Förderung und Selektion an und verankert sie per Gesetz. Weiter

[3] Das Grundgesetz regelt in Artikel 7 lediglich das folgende:

- Das gesamte Schulwesen steht unter staatlicher Aufsicht.
- Privatschulen sind unter bestimmten Einschränkungen zulässig.
- Religionsunterricht ist ordentliches Lehrfach; die Erziehungsberechtigten entscheiden über die Teilnahme des Schülers.
- Vorschulen bleiben aufgehoben.

heißt es:

„Art, Zahl, Umfang, Schwierigkeit und Gewichtung der Leistungsnachweise richten sich nach den Erfordernissen der jeweiligen Schulart und Jahrgangsstufe sowie der einzelnen Fächer. Die Art und Weise der Erhebung der Nachweise des Leistungsstandes ist den Schülern vorher bekannt zu geben; die Bewertung der Leistungen ist den Schülern mit Notenstufe und der Begründung für die Benotung zu eröffnen. Leistungsnachweise dienen der Leistungsbewertung und als Beratungsgrundlage."

Diese Ausführungen unterstreichen das Bedürfnis aller Beteiligten nach einer angemessenen Leistungsbeurteilung, die auch zur Förderung genutzt wird, und definieren darüber hinaus eine erste Handlungsanweisung: Die Beurteilungsmaßstäbe müssen transparent sein und den Schüler/-innen im Vorfeld einer Leistungsüberprüfung bekannt gegeben werden. Eine nachträgliche Festlegung der Korrekturmaßstäbe ist ausgeschlossen (womit eine weitere Elternbeschwerde aus Kapitel 1 geklärt wäre – sie war tatsächlich zutreffend). Der nächste wichtige Handlungsrahmen wird im Absatz 2 des Artikels 52 definiert:

„Die […] erbrachte[n] Leistung[en] werden nach folgenden Notenstufen bewertet:

sehr gut	=	*1 (Leistung entspricht den Anforderungen in besonderem Maße)*
gut	=	*2 (Leistung entspricht voll den Anforderungen)*
befriedigend	=	*3 (Leistung entspricht im Allgemeinen den Anforderungen)*
ausreichend	=	*4 (Leistung weist zwar Mängel auf, entspricht aber im Ganzen noch den Anforderungen)*
mangelhaft	=	*5 (Leistung entspricht nicht den Anforderungen, lässt jedoch erkennen, dass trotz deutlicher Verständnislücken die notwendigen Grundkenntnisse vorhanden sind)*
ungenügend	=	*6 (Leistung entspricht nicht den Anforderungen und lässt selbst die notwendigen Grundkenntnisse nicht erkennen)."*

Auf den ersten Blick wird deutlich, dass es an dieser Stelle um die bekannten Notenstufen geht. Auf den zweiten Blick ergeben sich jedoch noch weitere wichtige Anweisungen:

- Die Leistung wird an Anforderungen gemessen. *Anforderung* bezeichnet hier ein externes Kriterium, das aus dem Lehrplan herzuleiten ist. Die Leistung des Einzelnen darf also nur daran gemessen werden – und weder an der Klasse („Leistung ist deutlich besser als der Durchschnitt") noch an anderen Elementen („Leistung ist zwar gut, aber der Schüler zu unverschämt").
- Es ist von Anforderungen die Rede, und demnach muss es auch Kriterien geben, nach denen die Anforderungen eingeschätzt werden. In Kombination mit der oben genannten Forderung nach Transparenz wird dem Beurteiler damit geboten, für jeden Test im Vorfeld Kriterien zu erstellen, nach denen er die Anforderungen bewertet, und diese Kriterien den Schülerinnen und Schülern im Vorfeld bekannt zu machen. Die bereits angesprochene Vermutung, dass Lehrer/-innen die Bewertungsrichtlinien erst im Zuge der Korrektur schriftlicher Arbeiten festlegen oder bei einer mündlichen Prüfung die Leistung nach der eigenen Befindlichkeit einschätzen, ist also dezidiert gesetzlich untersagt. Dabei werden die Anforderungen nicht für jede Jahrgangsstufe landesweit festgeschrieben, sondern sie sind von der Lehrkraft aus den Zielen und Anforderungen des Lehrplans abzuleiten. Eine gegensätzliche Regelung hätte unerwünschte Konsequenzen, die in Kapitel 5 dieses Buches weiter ausgeführt werden.
- Die vier Abstufungen des Bestehens und die beiden Abstufungen des Nichtbestehens sind sprachlich

differenziert. Es ist also bereits die Grundlage für die Bewertung definiert: Der generelle Erwartungshorizont bei einer Beurteilung liegt in der allgemeinen Erfüllung der Anforderungen, nicht in der Maximalleistung. Darüber liegen zwei Abstufungen, die Leistungshandlungen beschreiben, in denen der zu Beurteilende die Anforderungen vollständig bzw. über alle Maßen gut erfüllt hat. Weist die Leistung bereits Lücken auf, entspricht aber im Großen und Ganzen noch den Anforderungen, wird die Notenstufe „ausreichend" rückgemeldet. In der Möglichkeit der sprachlichen Differenzierung der Leistungsstufen liegt übrigens auch der Hauptgrund für das bewährte Fünfer- bzw. Sechser-Notenschema: Hier ist es möglich, die Unterschiede zwischen den einzelnen Stufen sprachlich darzustellen. Bei einer 15- oder 100-stufigen Skala ist diese Unterscheidungsmöglichkeit nicht mehr gegeben, was die Differenzierbarkeit und damit auch die Begründbarkeit der einzelnen Notenstufen außerordentlich erschwert.

Zwei weitere wichtige Regelungen des Artikels 52 (BayEUG) definieren die Anforderungen an pädagogisches Handeln bzw. die Notwendigkeit einer professionellen Überprüfung der Leistung. Zum einen heißt es:

„Unter Berücksichtigung der einzelnen schriftlichen, mündlichen und praktischen Leistungen werden Zeugnisse erteilt. Hierbei werden die gesamten Leistungen eines Schülers unter Wahrung der Gleichbehandlung aller Schüler in pädagogischer Verantwortung der Lehrkraft bewertet."

Das bedeutet: Noten sind kein Ergebnis einer mathematischen Berechnung maschinell erhobener Messungen (vgl. zur Begründung hierzu Kapitel 5.4), sondern eine gewissenhaft durchgeführte Evaluation der Schülerleistungen durch einen Experten (vergleiche hierzu beispielsweise BayEUG, Art. 59, Abs. 1: *„Die Lehrkräfte tragen die unmittelbare pädagogische Verantwortung für den Unterricht und die Erziehung der Schülerinnen und Schüler"*). Dabei werden die pädagogische Verantwortung, der Grundsatz der Vollständigkeit („Unter Berücksichtigung *aller* Leistungen") sowie der Grundsatz der Gleichbehandlung besonders hervorgehoben. Letzterer definiert übrigens, neben der weiter unter beschriebenen Chancengleichheit, auch das wichtigste Kriterium für einen jeden Test: Er muss – so sagt man – gut sein. Welche Ansprüche bzw. Gütekriterien der Grundsatz der Gleichbehandlung an einen Test stellt, erfahren Sie in Kapitel 6.

Zum anderen wird definiert:

„Die Schulordnungen können vorsehen, dass in bestimmten Jahrgangsstufen der Grundschule und der Förderzentren, in Wahlfächern sowie bei ausländischen Schülerinnen und Schülern in Pflichtschulen und bei Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Pflichtschulen die Noten durch eine allgemeine Bewertung ersetzt werden."

Das heißt konkret: Besteht bei einem Lernenden ein besonderer Förderbedarf, kann der Beurteiler in Abstimmung mit der Schule den Lernenden aus der gebotenen anforderungsbezogenen Leistungsbeurteilung herausnehmen und individuell bewerten. Aus diesem Vorgang entsteht allerdings keine Note. Hierdurch verstößt die Lehrkraft auch nicht gegen den im Artikel 3 des Grundgesetzes bzw. Artikel 52 (3) des BayEUG festgelegten Gleichbehandlungsgrundsatz. Gleichbehandlung muss zwar die Grundlage jeglicher Leistungserhebung sein, aber gerade in diesem Fall kann man den Widerspruch zwischen Gleichbehandlung aller Schüler/innen und individueller Herausnahme einzelner förderbedürftiger Schülerinnen und Schüler mit einer positiven Lesart dieses Grundsatzes begründen. Danach müssen die Chancen für jeden einzelnen gleich sein, sodass bei gleicher Leistung die gleiche Note gegeben wird. Insbesondere schließt der Gleichbehandlungsgrundsatz bei der schulischen Leistungsbeurteilung Aspekte wie Willkür, Sympathie oder Standesdünkel aus. Diese Form der Gleichbehandlung definiert laut Aristoteles den qualitativen Charakter der Gleichheit: „Gerechtigkeit ist, Gleiche gleich zu behandeln – und Ungleiche ungleich". Das schließt logischerweise ein, dass wesentlich Ungleiches nicht gleich behandelt werden darf. Gerade bei Schülerinnen und Schülern mit besonderen Lernvoraussetzungen ist also eine Ungleichbehandlung, beispielsweise in Form eines Chancen- oder Nachteilsausgleiches, durchaus geboten:

Besonders benachteiligte Schülerinnen und Schüler erhalten bessere Noten, „als es die objektiven Leistungen rechtfertigen würden“ (Zielinski 1975, S. 882).

Damit hat der Lehrende als Prüfer einen gewissen pädagogischen Ermessensspielraum. Jürgens (2005) schreibt dazu:

„Weil jedoch aufgrund unterschiedlicher Eingangsvoraussetzungen die Schülerschaft einer Klasse heterogene fachliche und überfachliche Lernstände aufweist, würde eine durchgängige Anwendung von für alle Schülerinnen und Schüler gleichermaßen geltenden sachlichen Bezugsnormen keinen ‚Lern-‘ und somit auch keinen ‚Leistungs-Sinn‘ machen. Statt dessen scheint es geradezu geboten, unter Hinzuziehung individueller Bezugsnormen zieldifferente Anforderungen zu stellen, um möglichst vielen zu individuellen Lern- und Leistungserfolgen zu verhelfen, ohne die jahrgangsbezogenen bzw. allgemein geltenden Lehrplananforderungen allerdings aus dem Blick zu verlieren“ (S. 164).

Kurz gesagt hat die Lehrkraft also durchaus das Recht, Anforderungen zu modifizieren, solange sie es pädagogisch begründen kann.

Gleiches gilt beispielsweise auch bei längeren Erkrankungen eines Lernenden: Kann dieser längere Zeit seiner Pflicht, sich „so zu verhalten, dass die Aufgabe der Schule erfüllt und das Bildungsziel erreicht werden kann (BayEUG, Art. 56 (4))“ nicht nachkommen, greifen der Reihe nach unterschiedliche Regelungen, die von zusätzlichen Angeboten zum Nachholen des Stoffes (wobei allerdings die Eltern und der betroffene Lernende zuvorderst in die Pflicht genommen werden) bis hin zur oben beschriebenen individuellen Förderung reichen.

4.3. Regelungen der einzelnen Schulformen

Die nächste Hierarchiestufe des Schulrechts wird von den Verordnungen der einzelnen Schulformen definiert. In ihnen werden insbesondere die Bestimmungen zur Durchführung und Auswertung sowie die unterschiedlichen Formen von Leistungsüberprüfungen festgelegt.

Formen der Leistungsüberprüfung

Die Schulordnungen unterscheiden in der Regel zwischen
- mündlichen Leistungen (z. B. Ausfragen, Mitarbeit, Referate),
- schriftlichen Leistungen (z. B. Tests, Kurzarbeiten, Portfolios Fach- und Seminararbeiten) und
- sonstigen Leistungen (z. B. Projekte, Werkstücke).

Weiterführende Schulen definieren auch die so genannten
- Großen Leistungsnachweise (Schulaufgaben) und
- Kleinen Leistungsnachweise (Kurzarbeiten, Stegreifaufgaben, fachliche Leistungstests, Praktikumsberichte, Projekte sowie mündliche und praktische Leistungen), die in den unterschiedlichen Fächern mit unterschiedlichen Gewichtungen zur Zeugnisnote beitragen.

Hausaufgaben sind hingegen keine Leistungsnachweise. Sie werden nicht unter Aufsicht der Schule, also nicht unter prüfungsmäßigen Bedingungen angefertigt und sind daher grundsätzlich keine Leistungsnachweise im Sinne z. B. des Art. 52 Abs. 1 BayEUG (Bayerisches Erziehungs- und Unterrichtsgesetz). Etwas anderes gilt nur, soweit die Schulordnungen bestimmte zu Hause zu fertigende Arbeiten ausdrücklich bei der Auflistung der Leistungsnachweise nennen, so z. B. Seminararbeiten. Dies schließt eine indirekte Bewertung der Hausaufgaben über entsprechende Rechenschaftsberichte selbstverständlich nicht aus. So können Gegenstände, die zu Hause zu lernen waren, abgefragt und bewertet werden. Zusätzlich kann bei Hausaufgaben im Laufe eines Schulhalbjahres in Hinblick auf Regelmäßigkeit, Sorgfalt, äußere Form u. a. eine Wertung vorgenommen werden, die dann unter Wahrung des Grundsatzes der Gleichbehandlung aller Schülerinnen und Schüler in die Bemerkungen und Bewertungen über Anlagen, Mitarbeit und Verhalten der Schülerinnen und Schüler einfließt.

Darüber hinaus definieren die Schulordnungen die Regeln zur Ankündigung, zur Verteilung und zur Anzahl der unterschiedlichen Leistungstests für die einzelnen Fächer oder Fachgruppen. Zudem wird festgelegt, welche Hilfsmittel verwendet werden dürfen und ob eine Leistungsüberprüfung freiwillig zur Notenverbesserung wiederholt werden darf – was meistens nicht gestattet ist.

Durchführung der Leistungsüberprüfung

Neben den bereits in den Landesgesetzen getroffenen Aussagen zu den Gütekriterien werden in den Schulordnungen sehr wenige konkrete Hinweise zur Durchführung gegeben. Einige Aspekte bedürfen jedoch gesonderter Regelung:

- In der Regel werden in den einzelnen Schulordnungen die Anzahl sowie Form und Dauer der einzelnen Prüfungsleistungen für jedes Fach konkretisiert. Gegebenenfalls werden Bearbeitungszeiten und Korrekturfristen aufgeführt. Für manche Schulformen gibt es weitere Präzisierungen, beispielsweise dass kein großer Leistungsnachweis geschrieben werden kann, bevor ein vorheriger im gleichen Fach nicht korrigiert zurückgegeben und besprochen wurde (z. B. GSO Bayern § 57 (1)).
- Versäumt eine Schülerin oder ein Schüler die Durchführung eines Tests aus Gründen, die sie/er selbst zu verschulden hat, so kann der Test mit der schlechtesten Note bewertet werden. In welcher Form die Schülerin / der Schüler den Nachweis des Nichtverschuldens, z. B. bei einer Krankheit, zu erbringen hat, wird nicht weiter festgelegt und kann somit von der Schule selbst definiert werden. In der Regel wird jede Schule auf ein ärztliches Attest bei einem versäumten Leistungsnachweis bestehen, in diesem Fall ist der Schülerin / dem Schüler Gelegenheit zu geben, den Leistungsnachweis unter vergleichbaren Bedingungen zu erbringen (z. B. MSO Bayern § 47 (6)).
- Bedient sich eine Schülerin oder ein Schüler bei der Anfertigung einer zu benotenden schriftlichen oder praktischen Arbeit unerlaubter Hilfe, so kann die Arbeit ebenfalls mit der schlechtesten Note bewertet werden. Manche Schulformen definieren hier sogar einen Automatismus; Abschreiben führt also unweigerlich zur schlechtesten Note. Beim bloßen Versuch, zum Beispiel beim Bereithalten unerlaubter Hilfsmittel, kann ebenso verfahren werden (z. B. MSO Bayern § 47 (4)).

Interessant ist in diesem Zusammenhang folgende Fragestellung: Was passiert, wenn der Bewerter erst bei der Korrektur bemerkt, dass zwei Schüler/-innen absolut identische Arbeiten – inklusive Rechtschreibfehler – abgegeben haben? Kann man hier beide Schüler/-innen mit der schlechtesten Note bewerten, selbst wenn beide behaupten, die/der jeweils andere hätte abgeschrieben und man selbst habe nichts davon bemerkt?

Ob und wer in diesem Fall unerlaubte Hilfsmittel verwendet hat, kann vorerst nicht nachgewiesen werden. Derselbe Wortlaut in den Prüfungen beider Schüler/-innen deutet zwar sehr auf ein Abschreiben hin, allerdings liegen hierfür keine konkreten Beweise vor. So besteht zum Beispiel die Möglichkeit, dass sich die beiden Lernenden gemeinsam auf die Prüfung vorbereitet haben und folglich im Test eine auswendig gelernte Lösung niedergeschrieben haben.

Fraglich ist nun, wie der Lehrende in einer solchen Situation reagieren kann. Natürlich kann er versuchen, durch geschickte Gesprächsführung bei den Schüler/-innen zu ermitteln, welcher der beiden der Schuldige ist. Gelingt ihm das nicht, kann er die unerlaubte Hilfe (bzw. Hilfestellungstellung) zunächst nicht beweisen. Hier muss er auf den so genannten Anscheinsbeweis zurückgreifen (Cornelsen-Akademie 2011). Dieser besagt, dass man einen Zusammenhang annehmen darf, wenn er der normalen Lebenserfahrung entspricht. Dadurch wird die Beweislast umgekehrt, die beiden Schüler/-innen müssten glaubhaft beweisen, dass keiner von dem anderen abgeschrieben hat.

Aufgrund ihrer Expertise und Erfahrung kann die Lehrkraft bei vorliegendem Fall ein Verschulden beider Lernender annehmen. Bei einem Anscheinsbeweis würde vorausgesetzt, dass ein kausaler Zusammenhang zwischen den identischen Arbeiten und der Zusammenarbeit der Schüler/-innen besteht. In der deutschen Rechtsprechung wurde bei ähnlich gelagerten Fällen der Anscheinsbeweis als Methode der mittelbaren Beweisführung akzeptiert. Folgender Fall ist hierzu dokumentiert (Cornelsen-Akademie ebd.):

„Ein Prüfling lieferte eine Arbeit ab, die in Teilen wortwörtlich mit dem nur intern bekannten Lösungsmuster übereinstimmte. Darüber hinaus war die abgelieferte Arbeit auch in Gliederung und Gedankenführung mit der vorgesehenen Lösung identisch. Das Lösungsschema war nicht öffentlich erhältlich, sondern nur intern bekannt. Wegen einer erstaunlichen Übereinstimmung zweier Arbeiten entschied der Prüfungsausschuss auf einen Täuschungsversuch und schloss den Kandidaten von der Prüfung aus. Dieser klagte bis vor das Bundesverwaltungsgericht, weil er meinte, man könne ihm keine Täuschung nachweisen. Trotzdem unterlag er, weil das Gericht den sog. Anscheinsbeweis akzeptierte. Es folgte der logischen Schlussfolgerung des Prüfungsausschusses, der Prüfling habe nur über eine unerlaubte Täuschung zu diesen Informationen kommen können (BVerwG, Beschluss vom 20. 02 .1984). Für die Schule gibt es keine gesonderten gesetzlichen Regelungen; das Prinzip des Anscheinsbeweises gilt auch für Klassenarbeiten in der Schule, wenn die Übereinstimmung zwischen zwei Arbeiten so frappierend ist (z. B. Fehler an den gleichen Stellen), dass jede andere vernünftige Möglichkeit (gemeinsame Vorbereitung) ausscheidet" (Cornelsen-Akademie ebd.).

Das heißt: Der Lehrende darf annehmen, dass ein Prüfungsteilnehmer vom anderen abgeschrieben hat, wenn diese nicht glaubhaft nachweisen können, woher die erstaunlichen Übereinstimmungen kommen. Gelingt das nicht, kann die Lehrkraft bei ihrer Entscheidung bleiben, was auch von den Schulbehörden gedeckt wird. Im Zweifelsfall wird ein Gericht die Entscheidung treffen müssen.

Ob eine derartige Eskalation dieses Falles aber wirklich notwendig ist, erscheint fraglich. Getreu des schulischen Grundsatzes, dass die Erziehung und Förderung der Schüler/-innen im Mittelpunkt zu stehen hat, ist eine solch harte Klärung nicht zu empfehlen. Vielmehr sollte mit Erziehungsmitteln reagiert werden. Eine genaue Schilderung der Rechtslage und eine eindeutige Drohung sollten in diesem Fall genügen. Zudem könnte der Lehrende für die Zukunft präventive Maßnahmen ergreifen, um die Nutzung von Hilfsmitteln bei Schüler/-innen zu unterbinden: beispielsweise die beiden Kandidaten bei Prüfungen auseinandersetzen oder Trennwände zwischen den Schüler/-innen aufstellen, um ein mögliches Abschreiben zu verhindern. Konkret kann er auch die beiden betroffenen Schüler/-innen in die erste Reihe setzen, um sie bei zukünftigen Prüfungen besser im Blick zu haben.

Auswertung der Leistungsbewertung

Auch zur Auswertung selbst werden von den einzelnen Schulordnungen nur wenige Konkretisierungen vorgenommen, da das generelle Vorgehen bereits in den Landesverfassungen hinreichend definiert wird und an dieser Stelle nicht weiter präzisiert werden muss. Lediglich Fragen zum Zeitraum der Korrektur und zur Einsicht durch Schüler/-innen und Erziehungsberechtigte werden weiter ausgeführt (z. B. MSO Bayern § 46 (3); GSO Bayern § 57 (1)). Allerdings werden in diesem Zusammenhang auch Regeln für die nachträgliche Auf- oder Abwertung von Leistungstests durch eine nächsthöhere Instanz (z. B. Schulleitung) getroffen (z. B. LDO Bayern §27 (4); GSO Bayern § 54 (7)).

Hier gilt der Grundsatz, dass eine Bewertung die erbrachte Leistung wahrheitsgemäß wiedergeben muss. Damit besteht auch kein rechtlicher Anspruch auf Beibehaltung einer Note, die ersichtlich dem erbrachten Leistungsnachweis nicht entspricht. Vielmehr dürfen Noten von schriftlichen Prüfungen grundsätzlich nicht nur nachträglich verbessert, sondern auch verschlechtert werden, sofern für eine entsprechende Änderung ein sachlicher Grund gegeben ist. Dies ergibt sich indirekt z. B. aus § 27 (4) 2 LDO (Lehrerdienstordnung), wonach die Schulleiterin bzw. der Schulleiter im Einvernehmen mit der betreffenden Lehrkraft (oder bei entsprechendem Beschluss der Lehrerkonferenz) die Note einer schriftlichen Aufgabe ändern kann. Dies gilt übrigens auch für den Fall, dass die Erziehungsberechtigten die Note bereits zur Kenntnis genommen haben, da die Lehrkraft lediglich durch die bei der Leistungsbewertung zu beachtenden Grundsätze der Vollständigkeit, der Gleichbehandlung und der pädagogischen Verantwortung gebunden ist (StMUK, 2012). Inwiefern von dieser – rechtlich bestehenden – Möglichkeit Gebrauch gemacht wird, steht allerdings im pädagogischen Ermessen.

Dieser Paragraph der LDO beschreibt jedoch lediglich die Änderung bei strittigen Einzelnoten, nicht bei einer kompletten Arbeit. Die Notenänderung einer oder gegebenenfalls auch mehrerer einzelner Noten ist

also mit Einverständnis der Lehrkraft möglich, ohne Einigung obliegt die Entscheidung über eine Änderung der Lehrerkonferenz. §27 der LDO wird damit in aller Regel genutzt, um eine rechtlich korrekte Vorgehensweise nach einer Beschwerde oder bei umstrittenen Zeugnisnoten zur Verfügung zu haben. Eine generelle Abwertung eines Notenschlüssels, also die nachträgliche Auf- oder Abwertung aller Noten einer gesamten Arbeit, deckt diese Verordnung nicht ab.

Die von vielen Lehrkräften sehr gefürchtete nachträgliche Aufhebung einer Leistungsbewertung durch den Schulleiter bei nicht-angemessener Gestaltung eines Tests ist beispielsweise in §54 der Gymnasialen Schulordnung in Bayern definiert und lautet:

„Die Schulleiterin oder der Schulleiter kann nach Rücksprache mit der Lehrkraft und der Fachbetreuerin oder dem Fachbetreuer einen großen Leistungsnachweis für ungültig erklären und die Erhebung eines neuen anordnen, insbesondere wenn die Anforderungen für die Jahrgangsstufe nicht angemessen waren oder der Lehrstoff nicht genügend vorbereitet war."

Das bedeutet, dass die Schulleitung die Bewertung eines Tests aufheben und dessen Wiederholung anordnen darf, wenn dieser nicht nach den allgemein anerkannten und landesgesetzlich verankerten Gütekriterien erstellt wurde. Eine Wiederholung der Arbeit ist jedoch nur sinnvoll, wenn begründet festgestellt wird, dass der Test für die entsprechende Jahrgangsstufe nicht angemessen ist bzw. ob der Test nicht angemessen vorbereitet wurde (genaueres hierzu in Kapitel 6). Solange jedoch keine gerechtfertigte Begründung zur Annahme für Unangemessenheit vorliegt, ist beispielsweise das Argument des Notendurchschnitts („dieser Schnitt ist zu gut für die fünfte Klasse") oder der Notenverteilung irrelevant und damit nicht als Referenz für die Angemessenheit haltbar. Der Notendurchschnitt selbst sagt nämlich nicht gezwungenermaßen etwas über die Angemessenheit aus, da er durch zahlreiche Faktoren beeinflusst werden kann: Die Lernenden könnten beispielsweise außerordentlich fleißig oder generell leistungsstärker als vergleichbare Klassen gewesen sein – oder auch besonders unvorbereitet. Eine nachträgliche Änderung des Notenschlüssels für einen gesamten Test ist damit in der Regel ausgeschlossen, zumal das gegen das Gebot der Transparenz der Benotung verstößt. Auch an dieser Stelle sei nochmal auf die Episode 2 dieses Kapitels verwiesen.

Interessant ist in diesem Zusammenhang auch der so genannte Drittelerlass, ein Bestandteil einer schulischen Rechtsverordnung im Bundesland Nordrhein-Westfalen, gemäß derer bis 2006 eine Klassenarbeit vom Schulleiter genehmigt werden musste, wenn ein Drittel der Schüler/-innen kein ausreichendes Ergebnis erzielt hatte. § 6 Abs. (8) der APO SI (Ausbildungs- und Prüfungsordnung für die Sekundarstufe I des Landes Nordrhein-Westfalen) lautete bis zu seiner Aufhebung wie folgt:

„Erreicht bei einer Klassenarbeit ein Drittel der Schülerinnen und Schüler kein ausreichendes Ergebnis, entscheidet die Schulleiterin oder der Schulleiter nach Anhörung der Fachlehrerin oder des Fachlehrers, ob die Anforderungen angemessen waren. In diesem Fall ist die Arbeit zu werten, andernfalls ist sie zu wiederholen."

Als der Drittelerlass in NRW erstmals eingeführt wurde, stärkte er die Position der Lernenden gegenüber den Lehrer/-innen bei schriftlichen Leistungsnachweisen: Eine Klasse konnte beinahe konsequenzfrei die Leistungen in schriftlichen Tests verweigern, wenn der Test nicht ihren Vorstellungen entsprach. Angesichts der Aussicht, ihre Aufgabenstellung und den (Miss-)erfolg des vorherigen Unterrichts vor der Schulleitung rechtfertigen zu müssen, zogen es viele Lehrkräfte vor, lieber leichte Tests zu erstellen bzw. nachträglich die Leistungserwartung ihrer Klassenarbeiten herabzusetzen, sodass sich ein höherer Anteil ausreichender Leistungen ergab.

4.4. Regelungen auf Schulebene

In den Fällen, in denen Gesetze oder Verordnungen schweigen, sind zunächst die Schulen aufgefordert, Regeln über die Durchführung und Auswertung von Tests zu erstellen. Bekannte Fälle sind zum Beispiel

- die Attestpflicht bei versäumten Leistungsnachweisen,
- gesonderte Regelungen zur Auswertung (beispielsweise die Verpflichtung zur schriftlichen Niederlegung

eines Bewertungsschlüssels),
- die Verteilung der Tests über das Schuljahr, wobei die Verordnungen zu beachten sind,
- besondere Regelungen zur Korrekturfrist oder
- Regelungen zur Gegenzeichnung schriftlicher Arbeiten durch Erziehungsberechtigte.

Meist werden diese Regelungen von allen Schulen auf ähnliche Weise getroffen. Dort, wo die Schule keine Vereinbarungen trifft, liegt es in der Expertise des Lehrenden, die Leistungsüberprüfung sachgemäß durchzuführen. Dies gilt insbesondere für die Gestaltung von Testaufgaben (vgl. Episode 1). Hier hat jeder Bewertende einen hohen Gestaltungsspielraum. Dabei sollte man sich jedoch an den Grundsatz halten, dass ein Test die gleichen kognitiven Leistungen fordern soll, die im Unterricht von den Lernenden verlangt werden. Da diese Leistungen in der Regel vom Lehrplan bzw. den Kompetenzdefinitionen sehr ausführlich beschrieben werden, kann man sie in Prüfungen auch abfordern: Wenn beispielsweise Transferleistungen vorgeschriebene Anforderungen sind, sind diese selbstverständlich auch in Prüfungssituationen zulässig und geboten. Bei einer eventuellen Beschwerde gegen die Prüfung gilt also das Argument nicht, dass die betreffende Transferaufgabe nicht buchstabengetreu im Unterricht durchgenommen wurde. Es wird lediglich verlangt, dass der Transfer auch Teil des vorbereitenden Unterrichts sein muss. Das beinhaltet allerdings auch, dass man nicht einfach irgendetwas fragen und das dann als Transferaufgabe deklarieren kann – ein Bezug zum Unterricht und zum eigenen Fach muss herzustellen sein, und ähnliche Transferleistungen müssen im Unterricht bereits trainiert worden sein.

4.5. Beschwerden und Widersprüche

Da die Leistungsbeurteilung in der Schule nicht immer konfliktfrei verläuft, sollte man auch über die Möglichkeit der Beschwerde bzw. des Widerspruchs und die entsprechenden Verfahren Bescheid wissen. Wenn im Extremfall ein Lernender ankündigt, sie zu verklagen, ist es gut, die verschiedenen Verfahrenswege zu kennen.

Selbstverständlich gilt bei der Lösung von Unstimmigkeiten zwischen den verschiedenen Parteien auch im Rahmen der Leistungsbewertung der Grundsatz, dass zunächst einmal im Gespräch nach einer einvernehmlichen Lösung gesucht werden sollte. Ist dies nicht möglich, gibt es im Verwaltungsrecht festgelegte Regularien, nach denen gegen eine Entscheidung der Lehrkraft bezüglich einer Leistungsbeurteilung Beschwerde oder Widerspruch eingelegt werden kann.

Hierbei gilt als wichtigstes Kriterium die Unterscheidung zwischen dem Verwaltungsakt und dem Verwaltungshandeln. Der Verwaltungsakt ist die wesentlichste Handlungsform der öffentlichen Verwaltung und steht im Mittelpunkt der meisten Verwaltungsverfahren. Er ist vom schlichten Verwaltungshandeln zu unterscheiden, bei dem meist keine dezidierten Regelungen getroffen werden. Dabei gilt: Geklagt werden kann grundsätzlich nur gegen Verwaltungsakte, nicht gegen Verwaltungshandeln. Sie müssen demnach als Lehrender zwar in der Lage sein, jede Ihrer Entscheidung zu rechtfertigen, aber Sie können nicht für jede dieser Entscheidungen gleich verklagt werden.

Verwaltungsakte sind Entscheidungen, durch die die Schule ihre Einzelfallentscheidungen mit unmittelbarer Außenwirkung regelt. Unter Verwaltungsakten versteht man an bayerischen Schulen:
- die Aufnahme und Entlassung von Schüler/-innen,
- Versetzung und Nichtversetzung,
- Ordnungsmaßnahmen und
- Prüfungsentscheidungen (z. B. das Abiturzeugnis, Jahreszeugnisse, allerdings keine einzelnen Leistungsnachweise oder das Halbjahreszeugnis).

Für das Einspruchsverfahren gegen Entscheidungen der Schule ist diese Unterscheidung wichtig: Verwaltungsakte verlangen nach einem Widerspruch, Verwaltungshandlungen, wie zum Beispiel die Korrektur und Herausgabe eines einzelnen Tests, nach einer Beschwerde.

Widerspruch

Gegen Zeugnisnoten oder Nichtversetzungen kann man Widerspruch einlegen, da es sich bei diesen Entscheidungen um Teil eines Verwaltungsaktes (Zeugnis) handelt. Der Widerspruch kann innerhalb eines Monats nach Bekanntgabe des Verwaltungsaktes schriftlich oder mündlich zur Niederschrift sowohl bei der Schule, die den Verwaltungsakt erlassen hat, als auch bei der Schulaufsichtsbehörde eingelegt werden (§ 70 Abs. 1 Verwaltungsgerichtsordnung VwGO). Die Monatsfrist beginnt jedoch nur dann zu laufen, wenn der Betroffene schriftlich über den Rechtsbehelf des Widerspruchs belehrt worden ist. Ist diese Rechtsbehelfsbelehrung unterblieben oder unrichtig erteilt, beträgt die Widerspruchsfrist 1 Jahr (§§ 70 Abs. 2, 58 VwGO). Im Regelfall sollte man den Widerspruch an die betreffende Schule richten, diese ist verpflichtet, dem Verfahrensweg zu folgen.

Wenn die Schule den Widerspruch akzeptiert, hilft sie diesem ab, indem sie z. B. das Zeugnis bzw. die Entscheidung über die Nichtversetzung abändert. Wenn die Schule die im Widerspruch gegebene Begründung nicht akzeptiert, hilft sie ihm auch nicht ab und leitet ihn an die Schulaufsichtsbehörde weiter. Bei Realschulen, Gesamtschulen und Gymnasien ist das in Bayern die Bezirksregierung, bei Hauptschulen das zuständige Schulamt (kann das Schulamt dem Widerspruch nicht abhelfen, leitet es diesen weiter an die Bezirksregierung). Die Bezirksregierung entscheidet dann anhand des Widerspruches und den von der betreffenden Schule zu der Note oder der Versetzung schriftlich formulierten Begründungen, ob sie dem Widerspruch abhilft, also die Note ändert bzw. die Nichtversetzung rückgängig macht.

Der Widerspruch gegen einen Verwaltungsakt hat aufschiebende Wirkung (§80 VwGO). Aufschiebende Wirkung bedeutet, dass der Widerspruchsführer bis zum rechtskräftigen Abschluss des Verfahrens so zu behandeln ist, als wäre die angefochtene Entscheidung nicht ergangen. So kann beispielsweise eine Ordnungsmaßnahme vorläufig nicht vollzogen werden. Allerdings wird der Betroffene durch den Widerspruch in seiner ursprünglichen Rechtsstellung nicht verbessert; der nicht versetzte Lernende steigt daher nicht in die nächsthöhere Klasse auf und wer eine Prüfung nicht bestanden hat, kann nicht die Berechtigung in Anspruch nehmen, die sich aus dem erfolgreichen Abschluss ergibt. Weist die Schulaufsichtsbehörde den Widerspruch mit Rechtsmittelbelehrung und Begründung zurück, können die Betroffenen innerhalb eines Monats beim zuständigen Verwaltungsgericht Klage (Tresselt, 2013) erheben. Die Klage richtet sich dabei immer gegen das jeweilige Bundesland, da in Deutschland Schulen keine rechtsfähigen Anstalten des öffentlichen Rechts sind.

Beschwerde

Gegen einzelne Entscheidungen, z. B. gegen Einzelnoten oder Halbjahreszeugnisse, welche lediglich eine Rückmeldung bzw. Information der Eltern darstellen, kann kein Widerspruch eingelegt werden. Demnach sind auch keine Klagen vor einem Verwaltungsgericht möglich. Solche Angelegenheiten müssen im Rahmen eines Beschwerdeverfahrens beanstandet werden, falls der Versuch einer gütlichen Einigung, beispielsweise im Rahmen eines Elterngespräches, gescheitert ist.

Unter einer Beschwerde versteht man die Beanstandung eines bestimmten Verhaltens. Durch die Beschwerde soll es abgeändert werden oder es sollen Maßnahmen getroffen werden, um es in Zukunft zu verhindern.

Eine Beschwerde erfolgt bei der Schule oder wird, wenn sie dort erfolglos blieb bzw. wenn es sich um schulinterne Angelegenheiten handelt, über die Schule an die übergeordnete Schulaufsichtsbehörde weitergeleitet. Sie muss deutlich als Beschwerde gekennzeichnet sein; nicht jede Bitte um Überprüfung eines bestimmten Verhaltens ist eine Beschwerde. Außerdem muss sie begründet werden. Zu beachten ist, dass Beschwerden im Gegensatz zu Widersprüchen keine aufschiebende Wirkung haben (Tresselt, ebd.).

Ziel einer Beschwerde ist immer die Änderung einer Handlung oder eines generellen Verhaltens, also beispielsweise die Änderung einer als ungerecht empfundenen Note oder der Wunsch nach größerer Milde eines Bewerters. Sie zielt nicht auf Strafe oder Disziplinierung ab, wobei natürlich entsprechende Maßnahmen im Rahmen des Dienstrechts vorbehalten bleiben.

Wird der Beschwerde stattgegeben, benachrichtigt die Schulleitung die Beschwerdeführer. So kann die Schulaufsicht beispielsweise anordnen, dass eine Klausur erneut durchgeführt werden muss. Andernfalls

berichtet sie der Schulaufsichtsbehörde und erteilt dem Beschwerdeführer eine Abgabenachricht (ohne Begründung). Dagegen ist keine Klage mehr möglich. Für Einzelnoten kann ein Lehrer also nicht direkt verklagt werden.

Erfolgversprechend sind Beschwerden, Widersprüche und ggf. Klagen vor allem dann, wenn die Entscheidung der Schule nicht rechtmäßig und fachlich korrekt getroffen wurde. Dies kann der Fall sein

- aufgrund formaler Fehler, wenn z. B. nur drei Deutscharbeiten statt der verlangten vier geschrieben wurden.
- oder aufgrund inhaltlicher Fehler, also Korrekturfehler oder nicht-angemessene Tests.

Gerichte legen jedoch in der Regel den gesetzmäßig gegebenen Autonomiebereich des Lehrenden bei der Korrektur sehr großzügig aus. Sie nehmen bei einer Klage insbesondere nur den so genannten äußeren Rahmen, also die Rechtmäßigkeit der Maßnahme, unter die Lupe, das heißt: Sie prüfen, ob die Note aus Rand- und Schlussbemerkung begründet werden kann, urteilen aber nicht über die inhaltliche Passung zwischen Lösung und Korrektur.

Verstößt ein Lehrender im Rahmen der Korrektur wissentlich oder schuldhaft gegen die vorhandenen Bestimmungen, können weitere Sanktionsmaßnahmen folgen, die von der jeweiligen Entscheidungsebene (Schule, Schulaufsicht, Gericht) getroffen werden.

Zusammenfassend kann man im Hinblick auf die gesetzlichen Bestimmungen zur Leistungsbeurteilung an Schulen feststellen: So unkonkret manche Gesetze und Verordnungen auf den ersten Blick auch sind – sie enthalten höchst erfreuliche Konsequenzen für die professionell denkende Lehrkraft:

- Die Gesetze und Verordnungen weisen den Beurteiler an, sein Bewertungshandeln stets transparent und nachvollziehbar zu gestalten. Die bei Ingenkamp angesprochene Fragwürdigkeit der Zensurengebung wird dezidiert angesprochen; es werden Regelungen getroffen, die ein anfechtbares Verhalten untersagen.
- Die Kriterien für die Gestaltung des Anforderungsrahmens werden nicht festgeschrieben, sondern müssen aus den Lehrplänen abgeleitet werden. Hier hat der jeweilige Beurteiler einen hohen Gestaltungsspielraum, den er in pädagogischem Sinne nutzen soll. Der Gesetzgeber erkennt damit die Profession des Lehrberufs ausdrücklich an. Die Lehrkraft ist keine Prüfmaschine, sondern Pädagogin/Pädagoge.
- Willkür und eigene Befindlichkeit bleiben ausgeschlossen. Der Beurteilungsweg ist zu dokumentieren, die Note fachund sachgerecht zu begründen, damit bei eventuellen Beschwerden nachvollzogen werden kann, auf welchem Wege die Bewertung zustande gekommen ist. Dies dient in hohem Maße der Sicherheit, der Gerechtigkeit und der Gleichbehandlung.

4.6. Bilanzierung des Gelernten

In diesem Kapitel werden folgende Lernimpulse gesetzt:

- Bildung ist Ländersache. Damit bilden die Landesverfassungen die oberste rechtliche Instanz. Darunter folgen landesspezifische Gesetze, die schulartübergreifend und schulartspezifisch weitere Angelegenheiten regeln, sowie Regelungen auf Schulebene. Sie alle zusammen bilden den rechtlichen Rahmen, der das Lehrerhandeln bestimmt.
- Leistungsbeurteilung ist ein verpflichtender Bestandteil des Systems Schule. Ihre Ergebnisse sind in vorgeschriebenen und mündlich definierten Ziffernoten festzuhalten. Der Anforderungsrahmen hingegen ist von der Lehrkraft selbstständig aus dem Lehrplan abzuleiten, was ihr große pädagogische Freiheit bei der inhaltlichen Gestaltung seiner Prüfungen gibt.
- Einzelne Schüler/-innen können von der Benotung freigestellt werden, wenn es für eine angemessene Förderung geboten erscheint.
- Einzelne Noten können von Lehrkräften verändert, gesamte Arbeiten von Schulleitern zurückgenommen

werden. Allerdings gibt es für beides spezielle Regeln, eine willkürliche Zurücknahme einer Arbeit wegen allgemein zu schlechter Noten ist bspw. in aller Regel nicht gestattet.

- Schulische Handlungen fallen in den Bereich der Verwaltungsgerichtsbarkeit, in dem auf Verwaltungsakte (z. B. das Abitur oder die Nichtversetzung eines Lernenden) mit einem Widerspruch und evtl. einer Klage, auf Verwaltungshandeln (z. B. ein einzelner Test) mit einer Beschwerde reagiert werden kann.
- Ein Widerspruch geht den Dienstweg von der Schulleitung über die Schulaufsichtsbehörde bis hin zum Verwaltungsgericht; bei formalen Fehlern (z. B. weniger Leistungserhebungen als im Gesetz vorgesehen) ist eine Klage vielversprechend, ansonsten eher nicht.
- Eine Beschwerde hingegen geht nur bis zur Schulaufsicht und ist nicht einklagbar. Mit ihr soll ein Verhalten geändert oder unterbunden werden, z. B. eine zu strenge Benotung einer Klassenarbeit.

In Bezug auf die Episoden am Anfang des Kapitels bedeutet das Folgendes: Der Lehrer in Episode 1 hat eine Transferfrage gestellt, die aus dem behandelten Stoff ausgreift (neben dem fixen gespiegelten Ausschnitt geht es zusätzlich um den Blickwinkel des Betrachters). So eine Frage darf er auf Grund seines pädagogischen Ermessens auch dann stellen, wenn ihr Inhalt im Unterricht nicht explizit Thema war. Gerade um festzustellen, ob eine Leistung den Anforderungen nicht nur „voll", sondern auch „in besonderem Maße" entspricht (und damit die Note „sehr gut" erhält), können solche Fragen durchaus geeignet sein. Allerdings müssen die Schüler mit derartigen Transferaufgaben vertraut sein, und die Frage muss für die Schüler grundsätzlich zu beantworten sein. Ob das hier der Fall ist, kann ohne Hintergrundinformationen nicht endgültig festgestellt werden.

Die unglückliche Situation in Episode 2 enthält verschiedene Aspekte. Am eindeutigsten ist die Lage bei Elke: Nur weil Sie bei ihrer Mitschülerin einen Fehler übersehen haben, hat sie kein Recht, den Fehler ebenfalls nicht gewertet zu bekommen. Im Gegenteil: Sie müssten die Arbeit ihrer Freundin nach unten korrigieren. Dies ist aber nicht verpflichtend und pädagogisch in solch einer Situation durchaus fragwürdig. Anders sähe es aus, wenn Sie bei der Freundin eine Aussage als dezidiert korrekt markiert hätten, die Sie Elke angestrichen haben. Dann müssten Sie beides angleichen – nach oben oder nach unten. Dass den Schülern zu wenig Zeit zur Verfügung stand, ist ein formales Problem. Wenn das wirklich der Fall war, könnte es sein, dass die Arbeit auf Grundlage der für Sie geltenden Vorschriften formal ungültig ist. Ein nachträgliches Anheben der Noten ist dabei aber normalerweise ausgeschlossen, der Test müsste wiederholt werden. Ein schlechter Notenschnitt führt hingegen nicht automatisch zur Ungültigkeit der Arbeit.

Für die Episode 3 ergeht an dieser Stelle der Hinweis auf die geforderte Transparenz der Notengebung in Kombination mit sachbezogenen Kriterien bei der Leistungsbewertung. Damit kann der Vorwurf einer Korrektur aus eigener Befindlichkeit trefflich aus der Welt geschaffen werden. Der Nachweis von Transparenz und kriteriengestützter Korrektur ist für diesen Fall die einzig angemessene Reaktion. Dies wird im nächsten Kapitel weiter ausgeführt.

Die nächsten Kapitel thematisieren den Gestaltungsspielraum der Lehrkraft bei der Definition der Anforderungen und der Auswertung von Prüfungen. Auch hier gelten Regelungen, die nicht gesetzlich festgelegt werden müssen, da diese aus allgemein anerkannten wissenschaftlichen Standards entwickelt werden können und somit die Grundlage einer professionellen Bewertung sein müssen.

5. Messtheoretische Grundlagen: Bezugsnormen und Skalenniveaus

In diesem Kapitel werden die theoretischen Grundlagen des Messens von Leistungen beschrieben. Die Grundlage dieser Theorie ist ein Vorgang, den wir im alltäglichen Prozedere der Leistungsbeurteilung eher unbewusst vollziehen: wir messen. Welche Aspekte diesem Handeln methodisch zugrunde liegen und wie sich dieser Vorgang und seine Konsequenzen begründen lassen, ist Thema der nächsten Seiten. Nicht zuletzt erklärt dieses Kapitel auch die Berechtigung und die Grundlagen des Beurteilungshandelns.
Einige Aspekte dieses und des nächsten Kapitels entstammen aus dem Bereich der empirischen Sozialforschung, deren Forschungsmethoden und Begrifflichkeiten denen der Leistungsbeurteilung in der Schule sehr ähnlich sind. Auch in der empirischen Sozialforschung geht es darum, Dinge zu messen, die nicht direkt abgelesen werden können, sondern aus Handlungen und Aussagen geschlossen werden müssen. Auch die empirische Sozialforschung ist an größtmöglicher Messgenauigkeit interessiert und dient daher als methodische Basis für die zahlen- und messtheoretische Grundlage der Leistungsbeurteilung.
Hierfür müssen zunächst drei Fragen geklärt werden:

- Was bedeutet der Begriff „Messen" für den schulischen Bereich?
- Wie definiert man einen Vergleichsmaßstab für die Bewertung?
- Welche gültigen Gesamtaussagen lassen sich aus den gesammelten Noten herauslesen?

Diese Problemstellungen sollen anhand folgender Episoden illustriert werden:

Episode 1: Immer noch zu viele Fehler
Trotz zweier Grundschuljahre sind die Rechtschreibkenntnisse von Anna ungenügend. So macht Anna im ersten Diktat des Deutschunterrichts der dritten Klasse immer noch insgesamt 60 Fehler. Allmählich werden die Eltern unruhig, denn Anna soll nach der vierten Jahrgangsstufe aufs Gymnasium wechseln. Also werden verstärkte Anstrengungen unternommen, um Annas Rechtschreibfähigkeiten zu verbessern. Nach vielen Nachhilfe- und Übungsstunden ist eine positive Veränderung sichtbar: Im nächsten Diktat macht Anna nur 30 Fehler. Leider genügt diese Leistung immer noch nicht; die Note 5 beginnt bei 25 Fehlern – und für eine ausreichende Leistung sind sogar weniger als 15 Fehler notwendig. Es bleibt also trotz der Verbesserung der Rechtschreibfähigkeiten bei der schlechtesten Note.

Episode 2: Vergleichbarkeit von Leistungen
Die Rektoren der Volksschulen des Landkreises Auling beschließen: Um allen Schüler/-innen unserer Schule die gleichen Bedingungen für ihre Bewerbungen um Praktika und Arbeitsplätze zu gewährleisten, werden neben den – ohnehin standardisierten – Abschlussprüfungen auch sämtliche Probearbeiten der Jahrgangsstufen 8, 9 und 10 in allen Kernfächern als schulübergreifende Tests durchgeführt. Für jeden Tests gibt es standardisierte Korrekturbögen, die Verwendung von Multiple-Choice-Prüfungen zur weiteren Standardisierung der schulübergreifenden Leistungsmessung wird im Regelfall bevorzugt. Zudem werden die durchschnittlichen Ergebnisse jeder Schule veröffentlicht. Nach sechs Monaten ist vehemente Kritik der Eltern zu hören. Trotz der offensichtlichen Gleichbehandlung aller Lehrenden des Landkreises fühlen sich alle benachteiligt. Die Rektoren haben viele unbeabsichtigte Konsequenzen bei ihrem Beschluss übersehen. Welche?

Episode 3: Arithmetik über alles
Sie sind Mathematiklehrer einer 11. Klasse. Bereits einen Tag nach Herausgabe der Jahreszeugnisse steht der Vater eines Schülers vor dem Lehrerzimmer und wirft Ihnen vor, seinem Sohn aus purer Bosheit nur eine 2 im Zeugnis gegeben zu haben. Sie verweisen auf die Noten des Schülers: In den Schulaufgaben waren es je zweimal die Note 1 und 2, die mündlichen Leistungen waren 1, 1, 2, 3. Gemäß den schulischen Regeln werden Schulaufgaben doppelt gewichtet, so dass sich eine Endnote von 1,583 ergibt und damit leider knapp eine 2. Der Vater argumentiert jedoch ganz anders: Schulnoten sind ordinalskaliert. Bei ordinalskalierten Werten ist eine

Berechnung des Durchschnittswertes nicht zulässig; zumindest aber ergibt die Berechnung des arithmetischen Mittels eine Ungenauigkeit, die einen Wert wie 1,583 in den Grenzbereich zur 1 verschiebt und damit wie eine Zwischennote behandelt werden muss, was in seinen Augen die Zeugnisnote 1 ergibt. Wer hat nun Recht?

Diese Episoden illustrieren das Dilemma der schulischen Leistungsmessung sehr anschaulich: Jede Bewertung von Leistungen benötigt einen Vergleichsmaßstab, zu dem jeweils die individuelle Leistung in Relation gesetzt werden muss, um einen Wert oder zumindest ein Urteil für diese Leistung zu ermitteln. Die Erstellung bzw. Verwendung eines Maßstabs eröffnet Widersprüche, die nicht einfach zu lösen sind:

- Wie gehe ich damit um, wenn sich Schüler/-innen individuell verbessern, jedoch trotz dieser Verbesserung unter dem geforderten Niveau liegen?
- Was mache ich mit Schüler/-innen, die wegen körperlicher oder geistiger Voraussetzungen nicht in der Lage sind, die gleiche Leistung wie die anderen Schüler/-innen zu erbringen?
- Kann man noch von pädagogischer Diagnostik reden, wenn man alle Schüler/-innen ohne Berücksichtigung ihrer Voraussetzungen mit dem gleichen Maßstab bewertet? Muss man hier nicht differenzieren?
- Kurz gesagt: Wie findet man einen Mittelweg zwischen dem Gleichbehandlungsgrundsatz und der damit einhergehenden Ungerechtigkeit durch individuelle Leistungseinschränkungen?
- Hinzu kommt die Fragestellung nach der Bildung einer Gesamtnote aus den verschiedenen Einzelleistungen. Darf dabei einfach das arithmetische Mittel gezogen werden, oder sind weitere Verfahrensfragen zu bedenken?

5.1. Messen im schulischen Kontext

Jede Leistungsbeurteilung ist zunächst einmal eine Messung, und dieser Begriff verlangt nach einer genaueren Definition. Die Deutsche Industrienorm definiert Messen als „die Überführung von geplanten Tätigkeiten zu einer quantitativen Aussage über eine Messgröße durch Vergleich mit einer Einheit“ (DIN 1391-1).

Dieser in der Physik und den Ingenieurwissenschaften gewachsene Begriff wird oft auf andere Gebiete übertragen, wobei er allerdings mit einem anderen Sinn belegt wird. Für den Bereich der Leistungsbeurteilung liegt die besondere Schwierigkeit darin, dass festgesetzte Einheiten im Sinne einer technischen Messung nicht vorhanden sind: Leistung lässt sich nicht in Metern, Zoll oder Kilogramm messen. Leistungsmessung in der Schule ist damit zwar auch eine Zuordnung von Beobachtungen zu Maßeinheiten. Die Vergleichseinheit, also der Basismaßstab, muss bei der schulischen Leistungsmessung jedoch immer konstruiert werden. Dabei gelten folgende Regeln:

- Die Zuordnung muss struktur- und relationserhaltend sein, das heißt: Die Qualität der Leistung muss so abgebildet werden, dass der Maßstab ebenfalls Qualitätsaussagen zulässt. Man benötigt also eine entsprechende *Bezugsnorm* als Messgrundlage.
- Der Maßstab, also das Notenschema, muss diese Zuordnungsvorschrift stützen und damit ebenfalls Aussagen wie „besser“, „schlechter“ oder „gleich gut“ zulassen. Hierzu muss ein passendes *Skalenniveau* gewählt werden.
- Das methodische Vorgehen bei der Messung von Leistungen folgt dabei wissenschaftlichen Standards (in Form von *Gütekriterien*, siehe nächstes Kapitel), um den Einfluss des Messenden so weit wie möglich auszuschließen.

Im Grundsatz heißt das nichts anderes, als dass bei der Leistungsmessung so verfahren werden muss, dass die Beobachtung schlechter Leistungen durch als schlecht definierte Zahlen (=Noten) abgebildet wird, die Beobachtung guter Leistungen durch gute Noten. Die zugrundeliegende Messung muss immer ein methodisch möglichst gut beschriebener und nachzuvollziehender Vorgang sein, der prinzipiell von jedem geschulten Messenden durchgeführt werden kann. Dabei müssen alle Bewertungsvorgänge, also Beobachtung, Beschreibung, Interpretation und Bewertung, möglichst exakten Gütekriterien folgen, um Fehler so gut wie

möglich auszuschließen.

Für die beurteilende Lehrkraft ist es daher notwendig, alle testtheoretischen Schwierigkeiten und Unschärfen zu kennen, um das eigene Verfahren möglichst strukturiert, standardisiert und nachvollziehbar zu gestalten – auch gegenüber Dritten, z. B. Eltern oder Vorgesetzten. Die hierzu notwendigen Methoden sollen im Folgenden so dargestellt werden, dass die Planung und Durchführung der Leistungsbeurteilung an möglichst vielen theoretisch feststehenden Aspekten ansetzt und reflexiv evaluiert werden kann.

5.2. Bezugsnormen als Vergleichsmaßstab

Das Thema Bezugsnormen gehört, neben den Gütekriterien, zu den bekanntesten, aber auch zu den am häufigsten falsch interpretierten Themen der Leistungsbeurteilung. Bezugsnormen können zunächst weitgehend inhaltsleer definiert werden, zum Beispiel durch die Formulierung: *Unter einer Bezugsnorm versteht man einen Vergleichsmaßstab (also eine Norm), auf den sich ein Beurteilender bei der Messung einer Größe bezieht.*

Anhand dieser Norm kann eingeschätzt werden, ob die individuell zu messende Größe als vergleichsweise hoch, mittel oder niedrig beschrieben werden kann. Diese – einfach zu verstehende – Definition stößt dennoch eine intensive Auseinandersetzung für den schulischen Bereich an, da eine unsachgemäße Verwendung von Bezugsnormen die Qualität der Leistungsbeurteilung deutlich mindert.

Um aus dieser Definition Handlungsoptionen für die schulische Leistungsbeurteilung ableiten zu können, muss man die Bezugsnormen differenziert betrachten. Dazu werden im Folgenden die drei am häufigsten im schulischen Kontext genannten Bezugsnormen vorgestellt – die soziale, die *individuelle* und die *kriteriale (oder: sachliche)* Bezugsnorm (Rheinberg 2002):

- Die **soziale Bezugsnorm** erlaubt es, die Leistungen innerhalb einer definierten Gruppe miteinander zu vergleichen. Das bedeutet, dass die Testergebnisse in eine Rangreihe überführt werden, aus der hervorgeht, welcher Lernende bei einem Test weniger Fragen richtig beantwortet hat und wer besser abgeschnitten hat als seine Mitschüler/-in. Es geht dabei also nicht primär darum, inwiefern die Schülerin oder der Schüler die geforderten Kenntnisse und Kompetenzen gezeigt hat, sondern nur darum, ob er sie besser oder schlechter als die Mitschüler/-innen gezeigt hat. Außerhalb der Schule ist dieses Vorgehen für viele Problemstellungen schlicht notwendig: Immer dann, wenn aus einer kleinen, definierten Gruppe die besten (oder die schlechtesten) Gruppenmitglieder ermittelt werden müssen, kann die soziale Bezugsnorm zielführend angewendet werden. Dies kann mit folgenden Fällen illustriert werden:
  - Ein Arbeitsplatz soll besetzt werden. Aus der Gruppe der Bewerber/-innen soll die oder der Beste gefunden werden. Zweitrangig ist dabei, ob die Anwärterin oder der Anwärter die in der Stellenbeschreibung ausgeschriebenen Anforderungen erfüllt, denn eine Überqualifikation wird nicht gemessen.
  - Es gibt Gelder für die spezifische Förderung von fünf Schüler/-innen mit schlechten Lernleistungen (z. B. für gesonderten Nachhilfeunterricht). Auch hierbei ist es wichtig, die Personen auszuwählen, die die Förderung im Vergleich mit anderen am meisten benötigen, sodass die – meist knappen – Gelder zielführend verteilt werden können. Dabei geht es nicht darum, alle Kinder mit Förderbedarf festzustellen, sondern nur diejenigen, die im Vergleich zu den anderen jeweils den größten Förderbedarf haben.

  Die soziale Bezugsnorm ist in der Regel sehr einfach und unkompliziert anzuwenden. Würde man sie in schulischen Leistungsmessungen verwenden, gäbe es bei jedem Test eine gewissen Anzahl von Arbeiten mit der Note „sehr gut“ (denn einige sind immer die besten der Klasse), und auch eine gewisse Anzahl von Arbeiten mit der Note „ungenügend“ (denn einige sind immer die schlechtesten, und zwar unabhängig vom Wissenstand der Klasse insgesamt). Sie wird im Schulalltag immer noch manchmal verwendet, da sie einen scheinbaren Schutz vor zu gut oder zu schlecht ausfallenden Arbeiten verspricht. Es ist aber weder sachlich noch rechtlich korrekt, in Leistungsmessungen die soziale Bezugsnorm anzuwenden (Bügelmann et al. 2006). Zwar wird häufig kolportiert, dass Lehrkräfte ihre Korrekturen

damit vor Auffälligkeiten schützen, es gibt aber keine belastbare Studie über die Häufigkeit dieses Verfahrens, zumal es dezidiert nicht zulässig ist (vgl. Kap. 4).

- Die **individuelle Bezugsnorm** ist immer dann wichtig, wenn es darauf ankommt, der individuellen Entwicklung einen Wert zuzumessen. Hierbei spielen Gruppenvergleiche oder standardisierte Anforderungen keine Rolle, wichtig ist nur, ob sich die Leistung eines Individuums im Vergleich zu seinen früheren Leistungen gesteigert hat. Bei der Verwendung der individuellen Bezugsnorm handelt es sich also um ein längsschnittliches Verfahren, mit der der eigene Lernverlauf, aber auch die Schwankungen in der individuellen Lernerfolgskurve sehr gut nachvollzogen werden können. Diese Bezugsnorm ist gerade unter dem Aspekt der Förderung von Lernenden mit besonderem Bedarf wichtig. Durch die Ausrichtung des Feedbacks an den individuellen Aspekten und Leistungen kann sehr genau beschrieben werden, in welchen Teilbereichen ein Lernender besser oder schwächer geworden ist und worauf der Fokus bei der Planung der nachfolgenden Lernprozesse liegen sollte. Krampen untersuchte bereits 1987 die motivationalen Auswirkungen der individuellen Bezugsnorm und stellte fest, dass sowohl starke wie auch schwache Schüler/-innen davon profitieren – am meisten jedoch die Leistungsschwächeren. Eine professionell arbeitende Lehrkraft nutzt daher bei allen Schülerinnen und Schülern die individuelle Bezugsnorm in ihren Rand- und Schlussbemerkungen. Dort kann sie gesondert auf individuelle Fortschritte, besonderen Bedarf und erkennbare Verständnislücken eingehen, Fleiß und Anstrengung lobend hervorheben oder anmahnen und erkennen lassen, welche Lernhandlungen für die nächste Zeit im Fokus stehen sollten.
- Die **kriteriale, auf sachliche Kriterien ausgerichtete Bezugsnorm** ist der bekannteste der drei Vergleichsmaßstäbe. Den Maßstab für die Beurteilung bilden hierbei sachliche Kriterien, die für einen bestimmten Lern- oder Themenabschnitt, also z. B. für eine Unterrichtseinheit, erstellt werden. Die eigentliche Leistungsmessung gibt Auskunft darüber, auf welchem Niveau die gestellten Anforderungen erreicht wurden – und bereits diese Formulierung zeigt, dass der Gesetzgeber diese Bezugsnorm für die Schule hauptsächlich fordert (vgl. Kapitel 4). Diese Form der Leistungsmessung wird immer dann verwendet, wenn überprüft werden soll, ob ein Kandidat oder eine Kandidatin die Mindestkompetenzen für einen Beruf oder eine Tätigkeit besitzt (z. B. Jagdprüfung, Führerscheinprüfung). Sie ist immer dann wichtig und geboten, wenn die Ergebnisse der Leistungsmessung dazu beitragen sollen, Außenstehende (z. B. Ausbildungsbetriebe oder Universitäten) über den Stand einer Leistungsmessung zu informieren. Größtenteils werden die Testergebnisse in abgestufter Form mitgeteilt (z. B. bei Schulnoten oder Berufsabschlussprüfungen). Die jeweiligen Anforderungen werden vor der Prüfung aufgestellt, ein sozialer Vergleich ist bei der Leistungsbewertung nicht notwendig – so ist es beispielsweise bei der Führerscheinprüfung möglich, dass viele oder sogar fast alle die Mindestanforderungen erfüllen und damit die Zulassung zum Führen eines Kraftfahrzeugs auf öffentlichen Straßen erhalten.

Die Tabelle 1 fasst die Eigenschaften der drei Bezugsnormen kurz zusammen:

Bezugsnorm	Kennzeichen	Motivationale Auswirkungen (Krampen 1987)
Soziale Bezugsnorm	Die Leistungen des Einzelnen im Verhältnis zur Leistung einer Gruppe.	Bei leistungsschwächeren Schüler/-innen deutlich negativ, bei leistungsstärkeren neutral oder leicht positiv.
Individuelle Bezugsnorm	Individueller Lernfortschritt.	Bei leistungsschwächeren Schüler/-innen deutlich positiv, bei leistungsstärkeren positiv.
Sachliche Bezugsnorm	Sachliche Kriterien, unabhängig von der Gruppenleistung	Bei allen Schüler/-innen positiv, ohne dass eine bestimmte Leistungsgruppe deutlich von ihnen profitiert.

Tab. 1: Kennzeichen und motivationale Auswirkungen der drei Bezugsnormen

In der schulischen Praxis ist es ohnehin nicht geboten, eine der genannten Bezugsnormen in ihrer Reinform anzuwenden. Richtet man die Leistungsbewertung nicht nur an der Zuschreibung von Noten, sondern vor allem am Fördergedanken aus, kommt man um eine individuelle Bezugsnorm nicht herum (vgl. Band 1, Kapitel 6: Evaluationsphase). Zur Unterstützung von Lernprozess, Lernerfolg und Motivation der Lernenden ist es notwendig, dass der Einzelne eine präzise Rückmeldung darüber erhält, in welchen Teilbereichen er welches Lernverhalten und welche Form der Übung stärken könnte, um das zugrundeliegende Themengebiet tiefgehender zu erarbeiten. Es ist gerade unter Fördergesichtspunkten nicht zu vernachlässigen, dass auch die eigene Zufriedenheit mit der vorangegangenen Arbeit durch individuelles, auf die Bemühungen des Lernenden ausgerichtetes Feedback angestoßen werden kann.

5.3. Nachteile und unerwünschte Konsequenzen

In der Realität ist die Anwendung einer einzelnen Bezugsnorm in ihrer Reinform unmöglich, da es keinem Bewertendem gelingen wird, sich von Vergleichen, Kriterien oder individuellen Voraussetzungen frei zu machen. Eine alleinige Verwendung einer bestimmten Bezugsnorm ist darüber hinaus auch deswegen nicht zu empfehlen, da jede dieser Vergleichsmaßstäbe „blinde Flecken" (Rheinberg 2002, S. 64 ff.) aufweist, also bestimmte, für die Leistungsmessung wichtige Aspekte ausblendet. Aus diesem Grund verlangen Gesetze und Verordnungen, die einzelnen Bezugsnormen klar zu trennen. Den Kindern und Eltern ist ebenfalls klar zu machen, dass sich die Benotung der erbrachten Leistung auf die zu erreichenden Lernziele beziehen muss (kriteriale Bezugsnorm), dass Feedback und die Förderansätze jedoch auf die individuelle Leistungsentwicklung bezogen werden sollten (individuelle Bezugsnorm) und daher u. U. in ihrer Tendenz von der Gesamtnote abweichen können.

Blinde Flecken der sozialen Bezugsnorm

Je nachdem, ob sich Lernende in einer leistungsstarken oder leistungsschwachen Klasse befinden, erhalten sie bessere oder schlechtere Noten. So würde der gleiche mittelmäßige Schüler in einer guten Klasse untergehen, während er unter lauter schlechten Mitschülern glänzen könnte. Beides wäre nicht sinnvoll. Damit ist die soziale Bezugsnorm eher ungünstig, wenn Vergleichbarkeit zwischen Schulklassen, Schulen, Abschlüssen und anderen qualifizierenden Arbeiten gegeben sein soll.

Zudem bleibt Lernzuwachs aller Schüler/-innen unsichtbar. Da Noten immer im Vergleich, nie aber

abhängig von der Leistungssteigerung vergeben werden, werden individuelle Lernzuwächse, aber auch Schwankungen im Lernerfolg nicht oder nur verzerrt sichtbar: Jede Schülerin und jeder Schüler lernt dazu, aber solange sie oder er nicht mehr hinzulernt als seine Mitschüler/-innen, werden sie sich trotzdem nicht verbessern. Das kann sich gerade bei schwächeren Lernenden sehr ungünstig auf ihre Motivation auswirken.

Bei Anwendung der sozialen Bezugsnorm entsteht fast zwangsläufig eine Konkurrenzsituation unter den Schüler/-innen, denn letztendlich können Lernende ihre Noten nur verbessern, wenn sie andere Lernende überholen: Es wird immer nur ein paar gute Schüler geben.

Wenn ein Lernender im sozialen Vergleich stets die schlechteste Note erhält, kann man keine Aussage darüber machen, ob seine Leistungen aus sachlicher Sicht (und im Vergleich mit anderen Klassen) wirklich so schlecht sind, oder ob sie an sich gut sind, aber alle anderen Mitschüler einfach noch besser sind. Und wenn seine Noten sich verbessern, kann man keine Aussage darüber machen, ob seine Leistung wirklich aus sachlicher Sicht besser oder ob einfach ein anderer Teil der Klasse schlechter geworden ist.

Blinde Flecken der individuellen Bezugsnorm

Die individuelle Bezugsnorm blendet die Leistungsunterschiede zwischen Lernenden aus, was im Kontext der individuellen Förderung für leistungsschwächere Schüler/-innen motivierend sein kann. Allerdings können mit ihr überhaupt keine Vergleiche durchgeführt werden, weder mit den anderen Schülern, noch in Bezug darauf, wie sich der Lernstand des Einzelnen im Vergleich zu den Anforderungen verhält. Gerade das kann aber durchaus ein Anliegen von Lernenden sein. Insbesondere leistungsmotivierte Schülerinnen und Schüler sind daran interessiert, und auch für schwächere Schüler wäre es für die Selbsteinschätzung durchaus empfehlenswert, wenn sie nicht nur wüssten, wie viele Fortschritte sie gemacht haben, sondern auch, wie weit sie vom eigentlichen Niveau noch entfernt sind.

Sollte eine Bewertung ausschließlich aufgrund der individuellen Steigerung zustande kommen, würde es zudem für Schüler/-innen mit ohnehin schon überdurchschnittlich hohem Wissensstand immer schwieriger, durch noch weitere Steigerungen weiterhin gute Noten zu bekommen. Wer sich von einer schlechten auf eine mittlere Leistung steigert, erhielte demgegenüber relativ leicht eine gute Bewertung.

Aus diesem Grund wäre auch eine Vergleichbarkeit zwischen Qualifikationen und damit eine Zuweisung zu attraktiven Ausbildungs- und Arbeitsplätzen nicht mehr möglich.

Blinde Flecken der sachlichen Bezugsnorm

In ihrer Reinform angewendet besteht ein gravierender Mangel der sachlichen Bezugsnorm in ihrer sehr eingeschränkten Anwendbarkeit. Höherbewertete kognitive Leistungen wie Analyse-, Diskurs- oder Evaluationsfähigkeit können, strenggenommen, mit Hilfe der sachlichen Bezugsnorm nicht mehr getestet werden. Es ist trotz einer vorausgesetzten hohen diagnostischen Expertise des Bewerters schlichtweg nicht mehr möglich, für derartige kognitive Leistungen Sachkriterien aufzustellen, mit denen sich weitgehend interpretationsfrei beschreiben, interpretieren und bewerten lässt. Divergentes Denken und kreative Leistungen sind mit sachlichen Kriterien nur sehr schwer bis unmöglich zu testen.

Ein bekanntes Beispiel für unerwünschte Konsequenzen der sachlichen Bezugsnorm liefert das Amerikanische Schulsystem. Dort gibt es so genannte Hi-Stakes-Tests. Dabei handelt es sich um zentral erstellte (Abschluss-)arbeiten, die bei der Leistungszumessung stark an der sachbezogenen Bezugsnorm ausgerichtet sind und deren Ergebnisse schwerwiegende Konsequenzen für Schüler/-innen und Studierende haben: Sie entscheiden über das Vorrücken, darüber, welche Hochschule Sie besuchen können, und nicht zuletzt fließen die Ergebnisse der Tests auch in ein landesweites Ranking aller Schulen ein. Eltern und Schüler/-innen können also sehr einfach erkennen, welche Schule die besten Ergebnisse im zentralen Test hat.

Um die sachliche Bezugsnorm adäquat umsetzen zu können, bestehen diese Tests in der Regel aus Multiple-Choice-Befragungen, in denen primär Behaltens-, Verstehens- und Anwendungsleistungen, also konvergente

kognitive Leistungen, getestet werden können. Diese kognitiven Leistungen reichen jedoch in der Regel nicht aus, um Lernerfolg unter einem ganzheitlichen Bildungsaspekt zu beschreiben.

Hi-Stakes-Tests dienen offensichtlich nicht primär der Förderung, sondern der Zu- und Abweisung innerhalb eines selektiven Systems. Das hat weitreichende Konsequenzen: Da in den USA in der Regel keine Sprengelpflicht herrscht (bzw. die Sprengelpflicht über das Privatschulwesen umgangen werden kann), können Eltern weitgehend selbst entscheiden, auf welche Schule sie ihr Kind schicken. Eltern, die der Meinung sind, ihr Kind müsste später an einer der angesehenen Hochschulen studieren, wählen meist die Schulen, die im Ranking sehr weit oben stehen, sodass die Testergebnisse auch für Aspekte wie Klassenzusammensetzung und Lehrer/-innengehälter Bedeutung erlangen. Daraus resultierende hohe und niedrige Mittelzuweisungen und in letzter Konsequenz auch das Gehalt der Lehrkräfte hängen damit mittelbar von den Testergebnissen der Schule ab. Auch für den Universitätszugang sind die Konsequenzen eher ungemütlich:

„Entscheidend für die Universitätszulassung eines Studenten ist sein Abschneiden im sogenannten SAT (Scholastik Assessment Test). Von dem Druck der Studenten, im SAT gut abzuschneiden, lebt inzwischen eine kleine Industrie, die test-preparation companies. Princeton Review, die beste unter diesen Agenturen, bietet ein Deluxe-Paket für $ 7000 an; dem Käufer wird ein exzellentes Resultat garantiert" (Donoghue 2010, S. 632).

Um sie vergleichbar zu halten, werden in diesen Tests in erster Linie Aufgaben gestellt, die konvergentes Denken fördern und zu denen es Standardlösungen gibt, die als Anleitung für die Korrektur verwendet werden können. Meist werden 50–70 Multiple-Choice- bzw. Single-Choice-Fragen gestellt und zwei bis drei abschließende offene Fragen, für die knappe, eindeutige Antworten verlangt werden. Solche standardisierten Tests können kein kritisches Denken oder Schreiben, keine Problemlösekompetenz, Kreativität oder ähnliche wichtige Aspekte der Intelligenz erfassen. All das fällt somit aus den offiziellen Bewertungen eines Lernenden heraus.

Die Konsequenzen aus diesen Vorgaben sind relativ leicht abzuschätzen und werden von Sharon Nichols und David Berliner (2005) anschaulich dargestellt:

„Selbstverständlich sind die Tests und deren Standardlösungen bei Lehrern und Schülern bekannt und werden im Unterricht trainiert – das verlangen übereinstimmend Schulleitung, Eltern und Schüler. Problematisch wird diese Vorbereitung immer dann, wenn der Unterricht mehr auf das Format des Testes als auf die inhaltlichen Aspekte ausgerichtet wird und in der Schule damit Coaching statt Unterrichten stattfindet […]. Für die meisten Tests ist das verwendete Format der Aufgabenstellung und des Lösungsschemas im Vorfeld bekannt. Oft zirkulieren Aufgaben vergangener Runden im Lehrerkollegium oder werden sogar für Übungszwecke von der Schuladministration ausgegeben […]. Darüber hinaus bieten Schulbuchverlage und private Unternehmen eine Vielzahl von Artikeln zur detaillierten Testvorbereitung an" (S. 89).

Damit wird in letzter Konsequenz nicht der Unterrichtsinhalt, sondern der Test selbst zum Lernobjekt, was im Hinblick auf die Interpretationskraft der ermittelten Leistungsstände bzw. -steigerungen mehr als problematisch scheint. Nichols, Glass und Berliner (2006) fassen die mittlerweile sehr vielfältigen empirischen Ergebnisse in einem Satz zusammen: „To date there is no consistent evidence that high-stakes testing works to increase achievement." (S. 6). Die Folgen dieses hochstandardisierten und hochbedeutsamen Tests wurden bereits 1976 von Donald T. Campbell beschrieben (Campbells Gesetz): „The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (S. 85). Wie könnte diese Korruption des Systems aussehen? Hier nur einige Möglichkeiten:

- Die Verwendung unerlaubter Hilfsmittel durch Schüler/-innen und die Lehrperson wird wahrscheinlicher (Jacob/Levitt 2004).
- Tests sind nicht einfach nur Leistungsmessungen, sondern selbst Lernobjekt (Oelkers 2007).
- Leistungsschwache Schüler/-innen werden auf vielfache Weise daran gehindert, an den Hi-Stakes-Tests teilzunehmen, um den Schulschnitt nicht zu gefährden (Lind 2008).
- Lernschwache Kinder werden aus den verschiedensten Gründen (teilweise bewusst) benachteiligt.

- Hi-Stakes-Tests können intrinsische Motivation unterbinden (Krampen 1987; Deci/Ryan 2000).
- Das Schreiben eigenständiger Aufsätze wird zulasten standardisierter Lösungsvorschläge vernachlässigt.

Die Konsequenzen derartiger Tests für das schulische Handeln lassen sich sehr gut aus der Gegenüberstellung der Episode 2 und den beschriebenen Theorien ableiten: Im Endeffekt ist der Beschluss der Rektoren des Landkreises Auling unglücklich. Zwar existieren in der derzeitigen Diskussion eine Reihe von Standardisierungsbestrebungen, die meist mit dem Argument der Chancengleichheit durch Vergleichbarkeit der Leistungen begründet werden, die von den Rektoren gewählte Form bringt jedoch eine Reihe unerwünschter Konsequenzen mit sich:

- Da die Ergebnisse der Schulen veröffentlicht werden, werden sich Eltern „schwächerer" Schulen vermutlich darüber beschweren, dass ihr Kind an der (durch die Sprengelpflicht verpflichtenden) Schule nicht optimal gefördert wird und damit geringere Chancen auf dem Arbeitsmarkt hat.
- Eltern von Kindern auf stärkeren Schulen könnten die einseitige Ausrichtung auf die Testergebnisse, den damit entstehenden Druck und die Vernachlässigung höherer kognitiver Leistungen bemängeln.
- Auch die individuelle Förderung kommt vermutlich zu kurz, denn schließlich lassen sich schulische Testergebnisse auch dadurch steigern, dass man vor allem die guten Kinder fördert und bei den weniger guten nur noch dafür sorgt, dass sie nicht komplett versagen. Gerade wenige, herausragende Schüler/-innen beeinflussen, gemäß den Gesetzen der Mathematik, einen Durchschnittswert bedeutsam nach oben.

An dieser Stelle muss noch hinzugefügt werden, dass der Gesetzgeber die unerwünschten Konsequenzen einer sachlichen Bezugsnorm mit zentral bzw. landesweit streng ausdefinierten Kriterien anerkennt und die Lehrenden als Expert/-innen angemessen in die Pflicht nimmt. Mit feststehenden, verhältnismäßig interpretationsfreien Kriterien können lediglich Behaltens-, Verständnis- und einige Transferleistungen auch von unerfahrenen Bewertenden geprüft werden. Für höhere kognitive Leistungen ist dies für Novizen nicht mehr möglich und führt in letzter Konsequenz zum Ausblenden von nicht-testbaren Elementen wie Schlüsselqualifikationen, Kreativität oder Erziehungsaspekten. Zudem informiert eine rein sachorientierte Bezugsnorm streng genommen nur über die jeweils zum Testzeitpunkt beschriebenen Fertigkeiten oder Kenntnisse und blendet ebenso wie die soziale Bezugsnorm den Lernzuwachs aus.

Aus diesem Grund stellt der Gesetzgeber die diagnostische Expertise des Beurteilers weit in den Vordergrund. Alle Landesschulgesetze sind dabei sehr eindeutig und fordern im Regelfall von dem Bewertenden die Anwendung einer an curricularen Kriterien ausgerichteten Bezugsnorm. Bildungsstandards und Lehrpläne liefern einen geeigneten Rahmen; dabei sind Bildungsstandards kompetenz- und damit handlungsorientierter, was die Erstellung von Messkriterien deutlich vereinfacht. Aber auch die in den Bildungsstandards erweiterten kriterialen, d. h. sachorientierten Angaben zur Diagnostik bedürfen in der Anwendung hoher diagnostischer Expertise (vgl. Neuweg 2008). So heißt es beispielsweise in den Bildungsstandards für den Mittleren Bildungsabschluss im Fach Deutsch für den Kompetenzbereich Schreiben:

- „über Schreibfertigkeiten verfügen
- Texte in gut lesbarer handschriftlicher Form und in einem der Situation entsprechenden Tempo schreiben,
- Texte dem Zweck entsprechend und adressatengerecht gestalten,
- sinnvoll aufbauen und strukturieren: z. B. Blattaufteilung, Rand, Absätze,
- Textverarbeitungsprogramme und ihre Möglichkeiten nutzen: z. B. Formatierung, Präsentation,
- Formulare ausfüllen" (KMK 2006).

Illustriert werden diese Kompetenzen durch jeweils drei Kompetenzstufen und durch Beispielaufgaben. Doch auch diesen Definitionen ist nicht eindeutig zu entnehmen, für welche Leistung ein Lernender mit einer bestimmten Note belohnt werden soll. Aber das liegt auch nicht im Sinne der Bildungsstandards: Schließlich muss es darüber hinausgehend für den Lehrenden auch möglich sein, eigene Stärken und Schwerpunkte zu setzen und individuelles Feedback zur besonderen Rückmeldung hinsichtlich des

Lernprozesses einzelner Schüler/-innen zu geben.

Als Fazit kann man festhalten, dass der Gesetzgeber die Lehrkraft als professionell handelnden Experten für diagnostische und beurteilende Herausforderungen anerkennt und die Verantwortung für diesen schwierigen Prozess in ihre Hände legt. Dabei ist die Verwendung einer sachorientierten Bezugsnorm geboten, bei der die Kriterien von der Lehrkraft selbst aus dem Lehrplan abgeleitet werden. Sie sollte dabei mit der individuellen Bezugsnorm kombiniert werden, die als Feedback zum Anstoß von Lernprozessen (oder deren Veränderung) dient. Dadurch heben sich die blinden Flecken der kriterialen und der individuellen Bezugsnorm auf, ihre gemeinsamen Vorteile hingegen ermöglichen eine sachorientierte Bewertung und eröffnen Potenziale für Förderansätze.

5.4. Skalenniveaus

Nachdem die Frage nach dem zu verwendenden Basis- oder Vergleichsmaßstab geklärt ist, müssen als nächstes Regeln für die Zuordnung der Leistungen zu einer Note sowie zur Bildung einer Gesamtnote aus Einzelleistungen präzisiert werden. Hierzu sind Kenntnisse über Skalenniveaus erforderlich.

Zahlen können unterschiedliche Informationen kodieren, und je nach Art und Zustandekommen der Zahlen kann man mit ihnen unterschiedliche Rechenoperationen durchführen. Je nachdem, welche der dahinterliegenden Objekteigenschaften durch die Zahlen ausgedrückt werden können, unterscheidet man Skalen von unterschiedlichem Niveau. Die folgende Übersicht beschreibt die möglichen Skalenniveaus empirischer Messungen und nennt die Aussagen, die getroffen werden können (Bortz 1995):

Nominalskalierung. Nominalskalierte Werte haben lediglich die Funktion eines Etiketts und besitzen keine numerische Bedeutung. Zu den nominalen Daten gehören z. B. das Geschlecht oder ein Merkmal, welches vorhanden oder nicht-vorhanden sein kann (z. B. Brillenträger). Nominalskalen lassen nur Aussagen über die Gleichheit oder Ungleichheit der Messwerte zu und besitzen keine weiterführende numerische Bedeutung. Die einzigen Aussagen, die hier in Frage kommen, sind absolute oder relative Häufigkeitsaussagen. Wenn man z. B. bei einer Studie das Geschlecht der Teilnehmer mit 1 (männlich) und 2 (weiblich) kodiert, kann man zwar feststellen, dass es mehr weibliche als männliche Teilnehmer gibt, aber die Aussage, das Durchschnittsgeschlecht sei 1,7, ist sinnlos.

Ordinalskalierung. Ordinalskalierte Werte besitzen die Eigenschaften einer Nominalskala, geben aber zusätzlich die Rangreihe der Messungen wieder. Ordinale Daten können demnach der Größe/Wertigkeit nach geordnet werden und ermöglichen Aussagen im Sinne einer größer/kleiner/ gleich-Beziehung der Messwerte. Man spricht daher von einer Rangfolge. Allerdings sind die Abstände zwischen den Werten nicht unbedingt gleich groß, es werden sogar überhaupt keine Aussagen zu den Abständen verlangt. Es kann also sein, dass zwei Werte sehr dicht beieinander liegen, während zwei andere Werte einen großen Abstand zueinander haben. Daher ist es nicht möglich, Mittel- oder Durchschnittswerte zu berechnen (s. u.). Ordinalskalierung kommt z. B. bei Produktbefragungen vor, bei denen man die Zufriedenheit auf einer Skala von eins bis fünf bewerten soll. Hier kann man feststellen, welches Produkt bei einer Testperson den größten Anklang findet. Man kann aber nicht berechnen, um wie viel größer die Zufriedenheit im Vergleich zu den anderen Produkten ist. Ein weiteres Beispiel sind Rangkodierungen. Wenn man z. B. im Militär alle Ränge vom einfachen Gefreiten bis hin zum General durch Zahlen kodiert, kann man zwar feststellen, dass ein General höher steht als ein Hauptmann und dass es mehr Hauptmänner als Generäle gibt, aber die Aussage, der Durchschnittsrang sei 5,18, ist ebenso unsinnig wie die Aussage, drei Hauptmänner ergäben einen General.

Intervallskalierung. Intervallskalierte Werte besitzen die Eigenschaften nominaler und ordinaler Skalen. Zusätzlich können sie aber auch noch den Abstand zwischen den Messwerten aussagekräftig beziffern, denn Intervalldaten weisen immer den gleichen Abstand zwischen den einzelnen Werten auf. Diese Abstände werden als *Intervall* bezeichnet. Erst ab dem Intervallskalenniveau ist es mathematisch überhaupt möglich, ein arithmetisches Mittel zu berechnen. Ein Beispiel für eine intervallskalierte Skala ist die Celsius-Temperaturskala: Der Unterschied zwischen den einzelnen Graden ist immer identisch, und so kann man Durchschnittstemperaturen angeben. Die Aussage, 10 Grad Celsius seien doppelt so warm wie 5 Grad

Celsius, ist nicht möglich, da der Nullpunkt willkürlich festgelegt ist. Besser verständlich wird es, wenn man Uhrzeiten vergleicht. Die Aussage: „Jetzt ist es 17:00 Uhr, das ist doppelt so spät wie gestern um 8:30“ ist Unsinn.

Rationalskalierung. Eine Rational- oder Verhältnisskala besitzt alle Eigenschaften der Ordinal-, Nominal- und Intervallskalen. Zusätzlich dazu hat nun auch der Nullpunkt eine Bedeutung: *0* bedeutet hier *nichts*. In der Regel handelt es sich dabei um natürlich gegebene Zuordnungen wie die Körpergröße oder das Alter. Damit ermöglichen Rationalskalen nun auch Verhältnisaussagen, also Aussagen der Art „… doppelt so viel wie …“ oder „… doppelt so groß wie …“. Untersucht man z. B. die Altersstruktur der Bevölkerung, kann man selbstverständlich ein Durchschnittsalter angeben, und kann ebenso problemlos konstatieren, dass ein 40-jähriger doppelt so alt ist wie ein 20-jähriger.

Für die Betrachtung der Zuordnung einer Leistung zu einer Note ergibt sich damit eine zentrale Frage: Sind Schulnoten ordinal- oder intervallskaliert? Rein nominalskaliert sind sie nicht, da man sie ja ordnen kann (die Note eins ist besser als die Note zwei), und rational skaliert sind sie nicht, da man normalerweise keinen absoluten Nullpunkt angeben kann. Die Antwort auf diese Frage ist insofern wichtig, da die Jahresendnoten meist durch eine (gewichtete) Durchschnittsberechnung entstehen, ein Vorgehen, welches testtheoretisch erst ab dem Intervallskalenniveau erlaubt ist.

Falls Schulnoten intervallskaliert wären, müsste der Abstand von einer Schulnote zur nächstbesseren/-schlechteren einem präzisen und stabilen Leistungsunterschied zwischen Schüler/-innen entsprechen. Diese Annahme muss allerdings aus mehreren Gründen abgelehnt werden:

- Die Zuordnung der Leistungen zu einer Note ist nicht maßstabsgetreu bzw. mit Hilfe einer mathematischen Funktion möglich, da die zugrundeliegenden Leistungsanforderungen ebenfalls nicht in ein intervallskaliertes Niveau gebracht werden können. Als Beispiel: Anna hat in ihrer Mathematikprobe zwei Rechenfehler, Bert hat zwei Logikfehler. Diese Fehler können in keinem Falle als gleichwertig gelten, selbst wenn die Lehrkraft dafür bei beiden Lernenden je zwei Punkte abzieht.
- Selbst wenn es möglich wäre, Leistungsanforderungen intervallskaliert zu definieren, ist der an vielen Schulen gebräuchliche Zuordnungsschlüssel selbst nicht intervallskaliert. Die meisten Schulen haben sich darauf geeinigt, eine Leistung dann als ausreichend zu bewerten, wenn mindestens die Hälfte der Anforderungen erfüllt werden. Damit haben bereits die beiden Notenstufen 5 und 6 größere Intervalle als die vier Notenstufen, die den Grad des Bestehens klassifizieren. Illustriert werden kann dies am Beispiel des bekannten Bewertungsschlüssels der deutschen Industrie- und Handelskammern (0 bis 100 Punkte bzw. Prozentpunkte), der auch in der Schule häufig Verwendung findet:

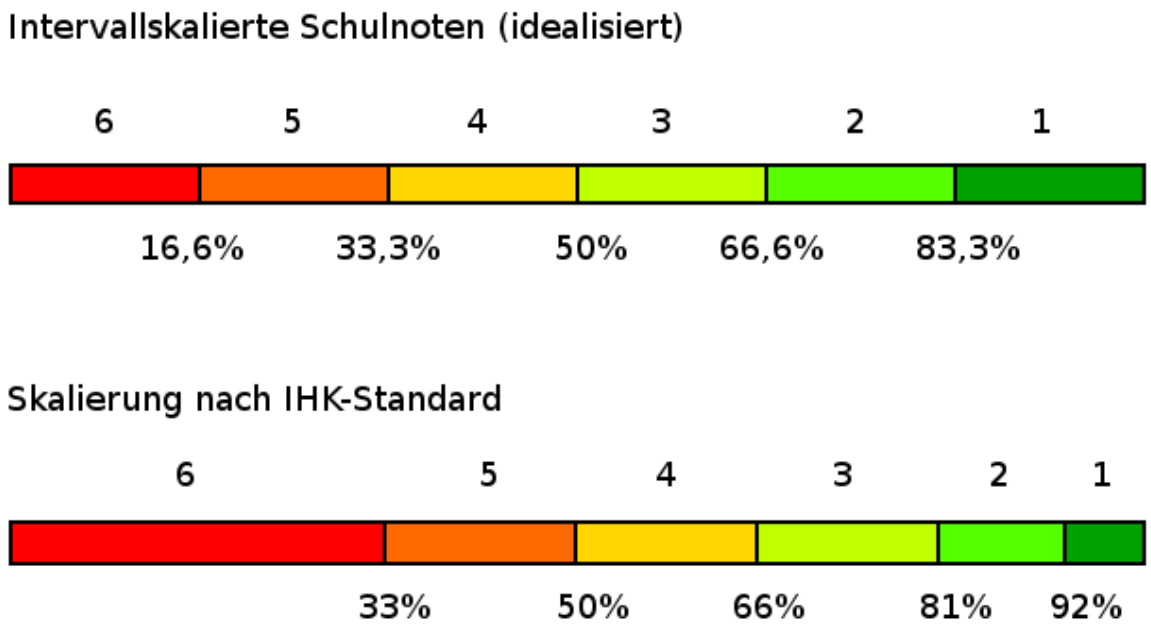

Abb. 21: Zwei unterschiedliche Zuweisungsschlüssel von Prozentpunkten zu einer Note

Es kann in der Praxis durchaus Konsequenzen nach sich ziehen, ob man die Durchschnittsnote in intervallskalierten Schulnoten oder mit Hilfe der Durchschnittspunkte nach IHK-Schlüssel berechnet. Als Beispiel seien hier die Einzelleistungen des Autors in der 11. Jahrgangsstufe Gymnasium in den Fächern Latein und Mathematik genannt. Die Verteilung der Noten entspricht den Tatsachen, die Einschätzung der Punkte nach dem IHK-Standard ist fiktiv. Es kann aber gut die Konsequenz für die Berechnung des arithmetischen Mittels illustriert werden, wenn alle Bewertungen im oberen bzw. unteren Bereich des Notenbereichs entstanden wären:

Fach: Latein	**Reale Noten**	**Fiktive IHK-Punkte**
1. Test	4	51 (= 4)
2. Test	5	34 (= 5)
3. Test	5	33 (= 5)
4. Test	3	67 (= 3)
Gesamtergebnis als Durchschnitt	4,25 (= 4)	46,25 (= 5)

Tab. 2: Vergleich der Durchschnittsnoten im Fach Latein mit unterschiedlichen Berechnungsgrundlagen. Die Schätzung der IHK-Punkte ist fiktiv, alle Punkte liegen jedoch im unteren Bereich des Zuweisungsraumes für die jeweilige Note laut IHK-Zuordnungstabelle.

Fach: Mathematik	Reale Noten	Fiktive IHK-Punkte
1. Test	1	99 (= 1)
2. Test	2	91 (= 2)
3. Test	2	91 (= 2)
4. Test	2	91 (= 2)
Gesamtergebnis als Durchschnitt	1,75 (= 2)	93,0 (= 1)

Tab. 3: Vergleich der Durchschnittsnoten im Fach Mathematik mit unterschiedlichen Berechnungsgrundlagen. Die Schätzung der IHK-Punkte ist fiktiv, alle Punkte liegen jedoch im oberen Bereich des Zuweisungsraumes für die jeweilige Note laut IHK-Zuordnungstabelle.

Je nach Berechnungsgrundlage ergeben sich in beiden Fächern unterschiedliche Noten – mal zum Vorteil, mal zum Nachteil des Autors. Dies gilt vor allem dann, wenn alle Einzelleistungen jeweils sehr knapp an den jeweils nächsthöheren bzw. nächstniedrigeren Grenzen liegen. Die Berechnung einer Durchschnittsnote enthält damit einen gewissen Unschärfewert. Problematisch wird es, wenn eine solche Verwendung ernste Konsequenzen hat, z. B. bei der Entscheidung über die Versetzung eines Lernenden.

Für den Kontext der schulischen Leistungsbewertung ergeben sich daraus wichtige Konsequenzen: Es gibt bei der Berechnung des arithmetischen Mittels aus ordinalskalierten Daten eine gewisse Unschärfe, deren Größe nicht bestimmt werden kann, weil die Methode der Zuweisung einer Note zu einer Leistung eben jene Unschärfe enthält. Man kann und sollte versuchen, durch genaues, exaktes und transparentes Arbeiten diese Ungenauigkeit möglichst klein zu halten; ganz eliminieren kann man diese Unschärfe jedoch nicht. Deshalb ist es durchaus eine Überlegung wert, ob eine Berechnung des Durchschnittswertes im Grenzbereich zwischen zwei Noten einen exakten mathematischen Schluss zulässt.

Andererseits wird durch die Ordinalskalierung der Daten die generelle Aussage eines arithmetischen Mittels nicht völlig ad absurdum geführt. Eine 2,6 kann zwar nicht ungeprüft in eine 3 überführt werden, sie ist aber mit Sicherheit keine 1 oder 4. Die Verwendung des arithmetischen Mittels bei einer Ordinalskala erzeugt zwar eine Unschärfe, welche durchaus einige Hundertstel oder Zehntel betragen kann, größere Ungenauigkeiten ergeben sich jedoch nicht. Damit ist die Diskussion vor allem in den Grenzbereichen zu führen.

Der Gesetzgeber berücksichtigt diese Ungenauigkeit. Beispielsweise wird im Bayerischen Gesetz über das Erziehungs- und Unterrichtswesen (BayEUG) Artikel 52 Absatz 3 gefordert:
„Unter Berücksichtigung der einzelnen schriftlichen, mündlichen und praktischen Leistungen werden Zeugnisse erteilt. Hierbei werden die gesamten Leistungen eines Schülers unter Wahrung der Gleichbehandlung aller Schüler in pädagogischer Verantwortung der Lehrkraft bewertet."

Eine gesetzliche Verpflichtung zur Bildung eines Durchschnittswertes gibt es also nicht; die Vergabe der Zeugnisnoten muss nur auf der Grundlage der erbrachten Leistungen erfolgen, dabei trägt der Lehrende die pädagogische Verantwortung. In der Regel sieht die einzelne Schule eine Verfahrensweise für diese Problemstellungen vor.

5.5. Bilanzierung des Gelernten

In diesem Kapitel werden folgende Lernimpulse gesetzt:

- Folgt man der gesetzlichen Basis, hat man grundsätzlich an Hand einer curricular definierten sachbezogenen Bezugsnorm zu beurteilen, d. h. nach vorgegebenen Kriterien, wie etwa den im Lehrplan verankerten Lernzielen.

- Dabei kommt der Lehrkraft als Expertin/Experten eine besondere Rolle zu, da die vorgegebenen Kriterien lediglich einen Rahmen abstecken. Die Auslegung und Umsetzung ist jedoch Aufgabe des Lehrenden.
- Allerdings hat jede Bezugsnorm für sich allein genommen blinde Flecken. Das gilt auch für die sachbezogene Bewertung, die ebenfalls unerwünschte Konsequenzen mit sich bringen kann und daher nicht in ihrer Reinform umgesetzt werden sollte.
- Die Anwendung der sozialen Bezugsnorm ist grundsätzlich unzulässig.
- In besonderen Fällen hat der Lehrende das Recht, Anforderungen individuell zu modifizieren, wenn er dies pädagogisch begründen kann. Dabei wird jedoch in der Regel der betreffende Lernende aus der regulären Bewertung herausgenommen und unter Verwendung der individuellen Bezugsnorm separat gefördert.
- Die Jahresendnote eines Faches wird unter Berücksichtigung aller Leistungen gebildet. Der Gebrauch des arithmetischen Mittels als Berechnungsgrundlage muss nicht schlichtweg abgelehnt werden, allerdings bedürfen Ergebnisse im Grenzbereich zwischen zwei Noten einer zusätzlichen Betrachtung.

Wenn Sie sich die Episoden zu Beginn des Kapitels vergegenwärtigen, sollten Sie die Probleme jetzt lösen können.

In Episode 1 wäre es sehr sinnvoll, den wirklich bemerkenswerten Leistungszuwachs von Anna im Rahmen der individuellen Bezugsnorm zu würdigen. Dies kann etwa durch Bemerkungen in oder nach der Arbeit, aber auch durch mündlichen Zuspruch geschehen. So bleibt ihre Motivation erhalten, sich weiter zu verbessern. Daneben kann man bei einer so schwachen Schülerin auch überlegen, Sie aus der Benotung herauszunehmen (vgl. Kap. 4). Das steht dem Ziel des gymnasialen Übertritts jedoch diametral entgegen. Das gleiche gilt natürlich auch für die Beschwerde des Vaters, der im Elterngespräch der ersten Episode des ersten Kapitels beklagt: „Mein Sohn hat sich doch so verbessert. Warum bekommt er immer noch die schlechteste Note?“

Dass es eine unglückliche Entscheidung der Aulinger Rektoren aus Episode 2 war, eine Art Hi-Stakes-Testing einzuführen, wurde bereits im Verlauf des Kapitels erläutert.

Bei Episode 3, in der sich ein Vater über die Errechnung der Endnote seines Sohnes beschwert, hat er mit seinen Vorwürfen grundsätzlich recht: Sie können aus den ordinalskalierten Noten keinen tragfähigen Durchschnitt errechnen, der so präzise ist, dass aus einer 1,583 automatisch die Note „gut“ wird. Aber das müssen Sie ja auch nicht: Soweit Sie nicht durch andere Regeln der Schule gebunden sind (auf die Sie dann auch verweisen könnten), entstehen die Durchschnittsnoten nicht automatisch, sondern aus Ihrer pädagogischen Verantwortung heraus. Wenn Sie also begründen können, dass die Noten insgesamt mehr auf eine „gute“ als auf eine „sehr gute“ Leistung schließen lassen, ist alles in Ordnung.

6. Kriterien für einen guten Test

Die in den letzten Kapiteln aufgeworfenen Fragestellungen führen zu der Herausforderung, Tests zur Leistungsbewertung so zu erstellen, dass mit ihnen gut gemessen werden kann. Tests müssen also gewissen Gütekriterien folgen. Was genau bedeutet das? Wann ist ein Test gut? Und vor allem: Wie kann bereits bei der Testerstellung dafür Sorge getragen werden, dass ein Test die Gütekriterien erfüllt?
Die aus diesen Fragen resultierenden Herausforderungen werden, wie gewohnt, mit Hilfe dreier Episoden illustriert:

Episode 1: Ein merkwürdiges Testergebnis
Die Auswertung Ihrer letzten schriftlichen Arbeit weist einige Merkwürdigkeiten auf. Zwar sieht die Verteilung der Noten auf den ersten Blick nicht sehr problematisch aus, dennoch haben Sie kein gutes Gefühl. Im Einzelnen erkennen Sie folgende Probleme:

- Aufgabe 1 wurde von keinem Lernenden richtig gelöst, von 5 Punkten wurden im Schnitt nur einer erreicht.
- Die Note 6 und die Note 1 kommen nicht vor.
- Alle guten Schüler/-innen konnten Aufgabe 3 nicht korrekt lösen, dafür haben fast alle schlechten Schüler/-innen bei dieser Aufgabe die volle Punktzahl.

Trügt Ihr schlechtes Gefühl? Welche Schlussfolgerungen ziehen Sie?

Episode 2: Spicken für alle
Sie vertreten eine erkrankte Kollegin bei der Aufsicht einer Probearbeit. Die Aufgaben selbst sind gut gestellt, allerdings stellen Sie während der Aufsicht fest, dass die Schüler/-innen eine etwas entspanntere Haltung zur Zusammenarbeit während eines Tests zeigen, als Sie es gewohnt sind. Kurz gesagt: Die Kinder spicken wie die Raben. Zur Rede gestellt behaupten die Schüler/-innen, dass dies bei ihrer Lehrerin durchaus üblich ist. Ihre Kollegin bestätigt dies. Sie meint: Gegen Spicken könne man eh nichts machen, und schließlich wird man nie alle unerlaubten Hilfsmittel finden. Darum kontrolliert sie generell nicht während einer Prüfung, damit haben alle Kinder die gleichen Voraussetzungen, was wiederum die Objektivität des Tests steigert. Sie zweifeln. Mit Recht?

Episode 3: Ein überraschendes Ergebnis
Herr Schütze, ein Englischlehrer, erstellt seine Prüfungsaufgaben schon seit Jahren nach dem gleichen Schema. Anders als in den Jahren zuvor fällt dieses Mal eine Schulaufgabe, die ebenfalls nach diesem Schema gestaltet ist, erstaunlich schlecht aus. Ein Drittel der Schüler/-innen erhält die Note Fünf, weitere 15 Schüler/-innen die Vier, vier Schüler/-innen die Drei, eine Schülerin die Zwei und keiner die Eins. Am Unterricht, so meint Herr Schütze, kann es nicht liegen, denn daran hat er nichts geändert. Er zieht in Erwägung, den im Vorfeld festgelegten Notenschlüssel im Nachhinein abzuändern, um sich nicht vor dem Direktor rechtfertigen zu müssen. Doch ist das die richtige Lösung?

Grundsätzlich gelten für die Testerstellung die bereits hergeleiteten gesetzlichen Vorgaben: Es müssen angemessene Tests zur Leistungsüberprüfung erstellt werden. Dies wird aus mehrerlei Gründen gefordert. Zum einen lässt ein guter Test die verlässlichsten Aussagen über den tatsächlichen Leistungsstand zu. Zum anderen unterstützt ein guter Test die pädagogische Verantwortung bei der Leistungsbeurteilung, da er eine differenzierte Begründung über das Zustandekommen einer Note zulässt.

Eine gute Ausgangsbasis zur Herleitung der Eigenschaften eines guten Tests bietet die klassische Testtheorie (Bortz 1995). Der Schwerpunkt dieses Modells liegt auf der Maximierung der Genauigkeit einer Messung bzw. auf der Minimierung des jeweiligen Messfehlers. Die klassische Testtheorie versucht zu klären, wie, ausgehend von einem Testwert einer Versuchsperson, auf die wahre Ausprägung der zu messenden Größe geschlossen werden kann. Dabei wird zunächst angenommen, dass das Testergebnis eigentlich dem „wahren" Wert des zu untersuchenden Merkmals entspricht, dass aber jede Messung oder jedes Testergebnis zusätzlich von einem Messfehler überlagert ist. Die Präzision eines Tests ist damit nur bestimmbar, wenn wahre Merkmalsausprägung und Fehleranteil getrennt zu ermitteln sind. Dies ist in der

Praxis jedoch nicht möglich. Daher muss versucht werden, den Messfehler möglichst weit zu reduzieren, um einen möglichst präzisen Test zu erhalten. Messfehler können in zufällige und systematische Anteile aufgeteilt werden.
Das bedeutet (Bortz ebd.):

- **Zufallsfehler** treten bei einer Messung zufällig auf. Sie sind nicht vorhersehbar oder beherrschbar, haben aber in der Praxis einen wesentlichen Vorteil. Durch das zufällige Auftreten sind sie unsystematisch, das heißt, sie sind nicht wiederholbar und gleichen sich über die Zeit hinweg aus. Im Kontext der Schule spricht man demnach dann von einem unsystematischen Fehler, wenn der unbeabsichtigt und zufällig passiert ist, wenn er nicht auf mangelhaftes Testverhalten oder Testgüte zurückzuführen ist und wenn er voraussichtlich in dieser Konstellation nie wieder passieren wird. Zufällige Fehler sind nicht zu kontrollieren, dies ist aber auch nicht notwendig, da diese auf lange Zeit gesehen durch Messwiederholungen keinen signifikanten Einfluss auf das Testergebnis haben. Ein Beispiel hierfür ist das zufällige Übersehen eines Rechtschreibfehlers. Solange solche Fehler nicht systematisch passieren, sondern auf einmalige Unkonzentriertheiten zurückzuführen sind, können sie durch genaues und präzises Messen verringert werden, sind aber für die Beurteilung der Testgüte nicht messbar und damit nicht relevant.
- **Systematische Fehler** hingegen verzerren das Testergebnis nachhaltig, da sie wiederholt und über einen längeren Zeitraum auftreten. Zudem haben sie im Gegensatz zum Zufallsfehler, bei dem das Ergebnis unvorhersagbar zu hoch oder zu niedrig ausfällt, eine Tendenz: Der Messwert wird für jeden Test gleichbleibend beeinflusst. Damit sind sie auch durch Wiederholungen nicht zu beseitigen. Das bedeutet aber auch: Kennt man die Quelle des systematischen Fehlers, kann er in der Regel eliminiert werden. Denn im Gegensatz zu Zufallsfehlern sind sie vorhersagbar und damit auch kontrollierbar. Typische Beispiele hierfür wurden in Kapitel 2 definiert. Sympathieeffekte sorgen beispielsweise dafür, dass die Noten eines Lernenden systematisch besser sind als die von seinen Mitschüler/-innen – trotz desselben Tests. Ursachen für einen systematischen Fehler bei einem Test können schlechte Messinstrumente oder die falsche Handhabung des Messinstruments sein.

Das bedeutet also: Je größer der systematische Messfehler, desto geringer ist der wahre Merkmalsanteil am Gesamtergebnis und desto weniger zuverlässig misst ein Test. Bei der Erstellung eines guten Tests muss demnach gewährleistet sein, dass das Ergebnis nicht durch systematische Fehler beeinflusst wird. Diese Maßnahmen zur Minimierung des systematischen Fehlers sind die Ausgangspunkte für die Kriterien eines guten Tests.

Konkret lassen sich sieben Kriterien identifizieren, an denen die Güte eines Tests festgemacht werden kann (Arnold 2002; Bortz/Döring 1995; Heller/Hany 2002; Weinert 2002). Diese Kriterien lauten:

Ein guter Test …

… ermittelt verteilte Ergebnisse, lässt also Qualitätsaussagen zu.
… hat einen angemessenen Schwierigkeitsgrad.
… trennt die guten von den schlechten Leistungen.
… ist gegeben, wenn jede Aufgabe ihren Teil zum Testergebnis beiträgt. Die einzelnen Aufgaben müssen also trennscharf sein.
… findet unter Rahmenbedingungen statt, die die Leistung nicht beeinflussen, ist also objektiv.
… stellt genaue und präzise Fragen, ist also reliabel.
… stellt Fragen, die das messen, was im Unterricht auch bearbeitet wurde, ist also valide.

Diese Kriterien werden im Folgenden weiter ausgeführt.

6.1. Verteilung der Ergebnisse

Durch eine Kombination von Durchschnitts- und Verteilungswerten kann man deutlich mehr über die Messwerte einer Gruppe aussagen als durch die alleinige Verwendung des arithmetischen Mittels. Bei der Testauswertung nach einer Korrektur fällt der erste Blick meistens, neben dem Notendurchschnitt, auf die

Verteilung der Ergebnisse, da man so den relativen Anteil der einzelnen Noten schnell erkennen kann.

Damit ist bereits das erste Gütekriterium – zumindest zum Teil – angesprochen: Ein guter Test muss Aussagen zu Qualitätsunterschieden zulassen, die Ergebnisse müssen also erkennen lassen, wie viele Schüler/-innen die Testaufgaben weniger gut und wie viele Schüler/-innen die Anforderungen besser gelöst haben. Unkorrekt wäre in diesem Zusammenhang ein Test, in welchem alle Schüler/-innen die gleiche Note bekommen.

Um dieses Gütekriterium besser zu definieren, muss an dieser Stelle auf eines der größten Missverständnisse zurückgegriffen werden: Es ist keine Anforderung an einen guten Test, dass die Ergebnisse normalverteilt sein müssen. In den Schulen wird manchmal noch die Meinung vertreten, dass Eigenschaften wie Intelligenz und Leistung normalverteilt seien und sich daher in jeder Klasse ein komplettes Spektrum von wenigen schlechten, vielen mittelmäßigen und wenigen guten Schülern finden müsse. Diese Verteilung müsse sich auch in den Testergebnissen einer Klasse wiederfinden: Ein guter Test solle demnach wenige schlechte, viele mittlere und wenige gute Noten produzieren. Diese Forderung ist aus den beschriebenen Gesetzen und Verordnungen allerdings nicht ableitbar. Die Widerlegung dieses Mythos ist sehr einfach, bedarf allerdings einiger grundlegender Kenntnisse über die Natur der Normalverteilung.

Bei der Normalverteilung handelt es sich nicht um eine einzige, bestimmte Verteilung, sondern um eine Familie von Verteilungen, die durch folgende Merkmale gekennzeichnet sind:

- Sie sind stetig (das heißt: stufenlos) und um den Mittelwert symmetrisch.
- Sie sind glockenförmig, das heißt: Die Mehrzahl der Werte ist um den Mittelwert gruppiert, weit abweichende Werte sind seltener.
- Für sehr große und sehr kleine Werte geht die Auftretenswahrscheinlichkeit gegen Null.
- Darüber hinaus ergibt sich eine Normalverteilungskurve durch die unendliche Wiederholung eines Zufallsexperiments. Viele Kenngrößen der Natur verteilen sich normal, zum Beispiel die Körpergröße von Frauen bzw. von Männern, die Fläche von Blättern, das Gewicht von Hummeln oder ähnliche natürliche Variablen.

Die Normalverteilung ergibt sich in der Natur immer dann, wenn die zu messende Kenngröße von einer Vielzahl von Faktoren abhängt, die jeweils unabhängig sind und zufällig variieren. Dies kann am Beispiel des Galton'schen Nagelbretts (Galton 1899, siehe Abbildung 6.2) anschaulich illustriert werden.

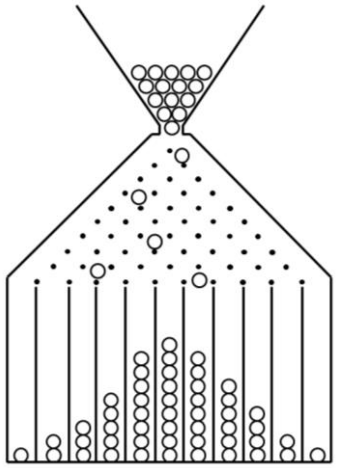

Abb. 24: Zweidimensionale Skizze des Galton'schen Nagelbretts. Jede Kugel fällt an einem Nagel mit 50%iger Wahrscheinlichkeit nach links oder nach rechts. Auf natürlichem Weg ergibt sich so eine Verteilung, die der Normalverteilungskurve ähnlich ist.

Überträgt man dieses Experiment ins Unendliche, ergibt sich die beschriebene stetige und symmetrische Normalverteilungskurve, wie sie in folgender Abbildung illustriert wird:

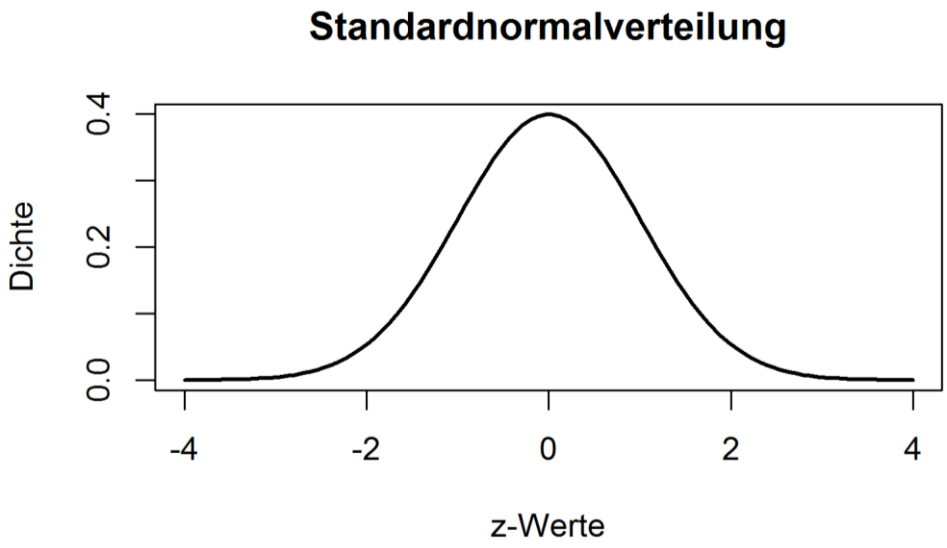

Abb. 25: Darstellung einer standardisierten Normalverteilungskurve

Mit Rückgriff auf den zentralen Grenzwertsatz, wonach zufällig gezogene Stichproben mit steigender Größe gegen eine Normalverteilung konvergieren, werden in der empirischen Sozialforschung üblicherweise Stichproben ab einer Größe von n > 80 als normalverteilt angenommen (Bortz/Döring 1995, S. 386). Die Population einer Schulklasse ist jedoch deutlich kleiner, da in einer Klasse selten mehr als 30 Schülerinnen und Schüler sitzen. Welche Auswirkungen dies für die Verteilungsannahme hat, kann folgendes Experiment illustrieren:

Es gibt einen statistischen Test, den so genannten Kolmogorow-Smirnow-Test, mit dem jede Verteilung daraufhin überprüft werden kann, ob sie der Normalverteilungsannahme folgt oder ob diese abgelehnt werden muss. Dieser Test funktioniert im Kern wie folgt:

- Man bestimmt eine Stichprobengröße und die Anzahl der Ausprägungen (also z. B. 25 Kinder in der Klasse und 6 Notenstufen).
- Nun wirft man für jede Person fünfmal hintereinander eine Münze. Die Anzahl der Köpfe (0–5) wird auf die möglichen Noten (1–6) transformiert und dieser Person zugewiesen. Wenn alle Personen eine Note haben, wird die Notenverteilung gezeichnet. Für die meisten dieser Experimente wird sich eine Kurve ergeben, die dem ersten Augenschein nach mit einer Normalverteilungskurve vergleichbar ist.
- Das Ganze macht man unendlich oft (in der Praxis erledigt das eine mathematische Formel).
- Nun ermittelt man die 5 % extremsten Verteilungen, also die Verteilungen, die am wenigsten mit der Normalverteilung zu tun haben.
- Wenn nun die Notenverteilung unserer Klasse extremer ist, als 95 % aller zufälligen Verteilungen, kann die Normalverteilungsannahme nicht mehr beibehalten werden.

Die folgende Grafik illustriert das Vorgehen: Auf der linken Seite der Grafik ist die Notenverteilung einer Schulklasse abgebildet. Die durchgezogene Linie illustriert die dazu passende stetige Normalverteilungskurve. Die rechte Seite trägt die Notenverteilung als Treppenfunktion ab, das heißt, die einzelnen Noten werden bei jedem Schritt aufsummiert. Gleiches geschieht auch mit der Normalverteilungskurve. Die gestrichelten Linien symbolisieren die 95 %-Grenzen der Normalverteilungsannahme. Wenn nun die Treppenfunktion unserer Notenverteilung eine der 95 %-Linien schneidet, ist die Notenverteilung nicht mehr als normalverteilt anzunehmen. In diesem ersten Beispiel ist dies nicht der Fall, man kann weiterhin von einer normalverteilten Kurve ausgehen.

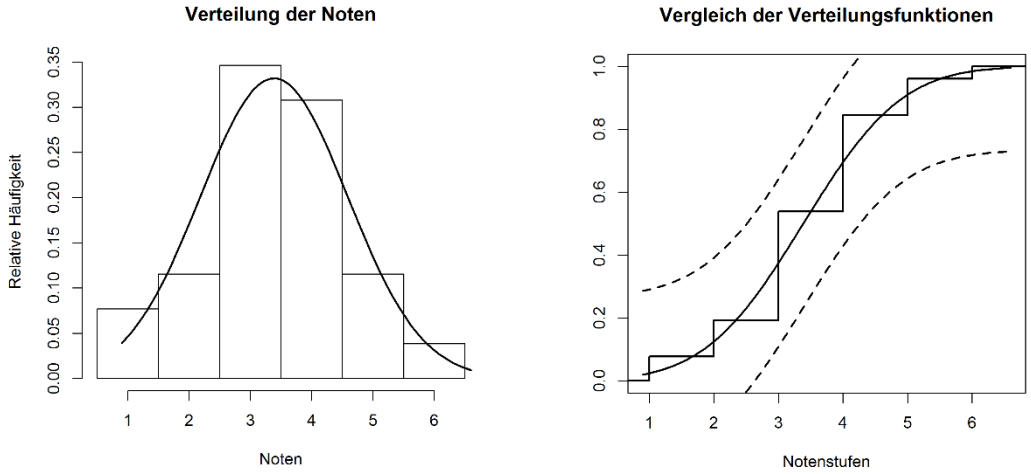

Abb. 26: Grafische Ausführung des Kolmogorow-Smirnow-Test für eine augenscheinlich unkritische Notenverteilung. Die Normalverteilungsannahme kann beibehalten werden.

Wie man am zweiten Beispiel gut erkennen kann, muss die Normalverteilungsannahme bei kleinen Gruppen auch bei extrem ungewöhnlich verteilten Einzelnoten nicht abgelehnt werden.

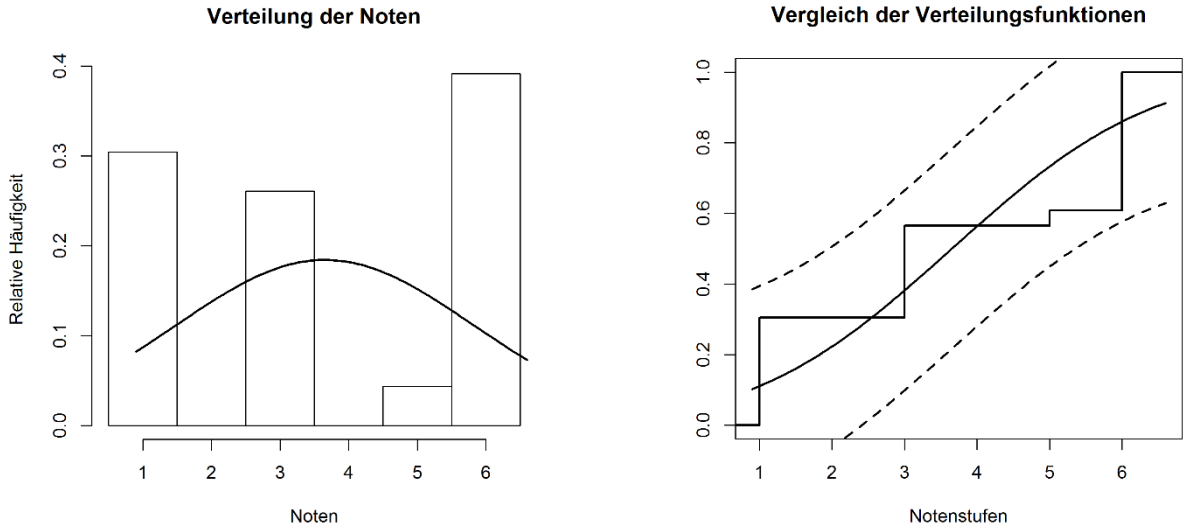

Abb. 27: Grafische Ausführung des Kolmogorow-Smirnow-Test für eine augenscheinlich kritische Notenverteilung. Die Normalverteilungsannahme kann dennoch beibehalten werden.

Bei dem diesen Zeilen zugrundeliegenden Versuch musste sehr lange nach einer Notenverteilung gesucht werden, die nicht mehr als normalverteilt angenommen wird. die folgende Abbildung zeigt das Ergebnis:

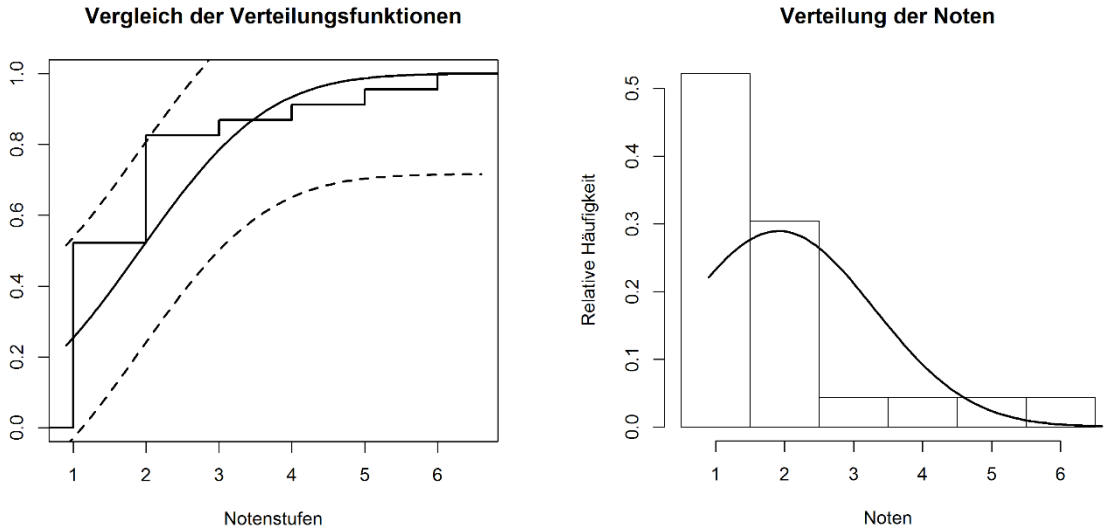

Abb. 28: Grafische Ausführung des Kolmogorow-Smirnow-Test für eine augenscheinlich kritische Notenverteilung. Die Normalverteilungsannahme muss abgelehnt werden, da die Verteilungskurve der Noten an der markierten Stelle den kritischen Grenzbereich schneidet.

Das bedeutet: Bei kleinen Gruppen, zum Beispiel bei einer Schulklasse, sind fast alle Notenverteilungen mathematisch als normalverteilt anzunehmen, selbst wenn die Inaugenscheinnahme ein anderes Ergebnis erbringt.

Es wäre also ein unangemessenes Vorgehen, die Notenverteilung erst nachträglich aufgrund der Kennzahlen des Leistungstests so festzulegen, dass dabei eine Normalverteilungskurve entsteht. Diese Erkenntnis ist nicht neu (vgl. Kapitel 4 und 5). Dennoch hat der Mythos der Normalverteilung von Testergebnissen einen wahren, testtheoretisch begründbaren Kern. Folgende Annahmen wurden bereits angesprochen:

- Die Normalverteilung ist eine natürliche Verteilung, die sich bei vielen Aspekten des menschlichen Lebens für große Gruppen einstellt.
- Ein guter Test bildet den wahren Wert möglichst präzise ab.

Nun hängt es also davon ab, wie die Leistung einer großen Gruppe verteilt ist. Die westliche Welt nimmt an, dass sich auch die Leistungsfähigkeit beim Menschen normal verteilt (Bortz/Döring 1995). Das heißt: Es gibt wenige sehr schlechte Leistungen, wenige sehr gute, die meisten Menschen zeigen eine mittlere Leistung. Wenn diese Annahme stimmt, dann müssen sich in vielen kleineren Gruppen bei Leistungstests ebenfalls Verteilungen ergeben, die zumindest nicht als normalverteilt abgelehnt werden müssen.

Ob diese Annahme der Normalverteilung von Leistungen stimmt, ist allerdings umstritten und würde außerdem dazu führen, dass es immer sehr schlechte Leistungen gibt, was aus Fördergesichtspunkten für Schulklassen keine geeignete Annahme sein kann: Gute Leistungen aller Lernenden durch guten Unterricht wären damit per definitionem unmöglich.

Darüberhinaus kann man nicht für jede natürliche Variable eine Normalverteilung voraussetzen. Gleiches gilt auch für den Schulkontext. Es gibt eine Reihe weiterer natürlicher Verteilungen, die ebenfalls häufig zu finden sind – beispielsweise die Pareto-Verteilung bei der Beteiligung von Schülerinnen und Schülern im Unterricht: Nur wenige Lernende sind besonders aktiv oder melden sich oft, viele Schüler/-innen sind eher unauffällig, wie es die folgende Abbildung illustriert.

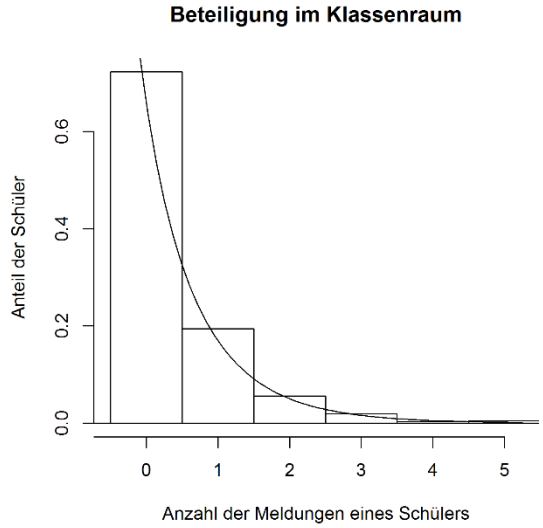

Abb. 29: Verteilungen der Meldungen im Unterricht während eines Schultages. Die Aktivitätsverteilung ähnelt einer Pareto-Verteilung.

Es gibt gute Gründe für die Annahme, dass sich auch Leistung umgekehrt paretoverteilt: Viele Menschen zeigen gute bis sehr gute Leistungen, schwächere Leistungen sind wenige zu erkennen. Beispielsweise geht die japanische Gesellschaft von einer umgekehrten Pareto-Verteilung der Leistungen aus, dementsprechend sehen auch die Notenverteilungen aus. Ob eine Normalverteilungsannahme bei Leistungstests eine Rolle spielt oder nicht, hängt also von der zentralen Annahme zur Verteilung des zugrundeliegenden Wertes, in diesem Fall der Leistung, ab.

Das erste Gütekriterium muss also lauten: Ein guter Test ermittelt verteilte Ergebnisse, lässt also Qualitätsaussagen zu. Er muss aber nicht zwangsweise normalverteilt sein.

6.2. Schwierigkeit

Schwierige Tests werden nur von wenigen Schüler/-innen richtig gelöst. Bei leichten Tests kommt dagegen fast jeder zum richtigen Ergebnis. Die Schwierigkeit eines Tests beeinflusst damit ganz wesentlich die Verteilung der Punkte- und damit der Notenwerte. Gleiches gilt sinngemäß für jede einzelne Aufgabe. Ein Test, und in ihm auch jede Aufgabe, sollte weder zu leicht noch zu schwer ausfallen: Ein Test, der von allen Schüler/-innen komplett gelöst wird, bringt ebenso wenig wie ein Test, an dem alle Lernenden scheitern.

In diesem Zusammenhang benötigt man eine Möglichkeit, den Vorwurf eines zu schweren oder zu leichten Tests entkräften zu können, da sich aus der Schwierigkeit auch Rückschlüsse auf die Angemessenheit eines Tests ziehen lassen.

Für das Kriterium der Schwierigkeit wird im Folgenden wiederum auf die Erkenntnisse der empirischen Sozialforschung zurückgegriffen, denn auch bei Befragungen ist es ein Anliegen, dass die gestellten Fragen weder von keinem einzigen noch von allen Teilnehmern beantwortet werden können. Das ergäbe unschöne Decken- oder Bodeneffekte, mit denen im Folgenden keine weitergehenden Aussagen getroffen werden könnten. Es gilt folgende Regel zur Berechnung der Schwierigkeit einer Aufgabe (Bortz/Döring 1995):

- Bei dichotomen Fragen (Ja/Nein-Fragen) ist der Schwierigkeitsindex durch den Anteil der Personen beschrieben, die die Fragestellung bejaht oder richtig gelöst haben.
- Bei mehrstufigen Fragen oder bei Fragen, auf die es mehr als einen Punkt gibt, lässt sich der Schwierigkeitsindex mit der Formel *SI = Durchschnittlich erreichte Punkte/Maximalpunktzahl* berechnen.

In jedem Fall kommt dabei ein Wert zwischen 0 und 1, also zwischen 0 % und 100 % heraus. Ein Schwierigkeitsindex von 0 kennzeichnet damit eine Frage, die kein Lernender richtig gelöst hat, der Wert 1 besagt, dass alle Befragten diese Aufgabe korrekt gelöst haben. Eine leichte Frage hat also einen hohen

Schwierigkeitsindex, eine schwere Frage einen niedrigen.

In der empirischen Forschung geht man davon aus, dass die Schwierigkeit jeder Frage und des kompletten Tests nicht unter 0,2 und nicht über 0,8 liegen darf, da die Aufgabe bzw. der Test ansonsten zu schwer oder zu leicht ist. Auf die Schule übertragen lautet dies: Im Durchschnitt müssen die Schüler/-innen für jede Aufgabe mindestens 20 % der möglichen Punkte erreichen, da diese Aufgabe ansonsten zu schwer ist. Wenn der Durchschnittswert der erreichten Aufgabe 80 % der möglichen Punkte übersteigt, ist die Aufgabe zu leicht.

Es genügt, den Schwierigkeitsindex für jede Aufgabe zu berechnen. Wenn die Schwierigkeitsindizes aller Aufgaben im erwünschten Korridor liegen, ist der Test automatisch weder zu schwer noch zu leicht.

6.3. Streuung

Es erscheint naheliegend, dass vor allem mittelschwere Fragen gestellt werden sollten, um zuverlässig ein mittleres Schwierigkeitsniveau zu erreichen. Dieser Schluss ist jedoch nicht richtig. Bei der Erstellung eines Test ist nämlich ein weiteres Testgütekriterium zu bedenken: Der Test muss streuen, er muss es also ermöglichen, diejenigen Schüler/-innen, die bei diesem Test eine schlechte Leistungen zeigen, von denjenigen zu trennen, die gute Leistungen zeigen. Zudem muss er erlauben, auch sehr gute und sehr schlechte Schüler/-innen zu identifizieren, er muss also im Idealfall das komplette Spektrum der möglichen Aussagen nutzen. Ein Test, der nur aus mittelschweren Fragen besteht, wird zwar schlechte von guten Schüler/-innen trennen, sonst aber kaum weitere Aussagen zulassen. Ein guter Test muss also der Tendenz zur Mitte entgegenwirken.

Dieses Kriterium zu erfüllen klingt einfacher, als es in der Praxis ist. Denn man darf die Noten ja nicht nachträglich festlegen, sondern muss die Auswertung des Tests bereits im Vorfeld transparent machen. Hier helfen, neben der Erfahrung, folgende Faustregeln:

- Die Ergebnisse eines Tests werden streuen, wenn er eine Mischung aus schweren, mittleren und leichten Aufgaben enthält (allerdings keine zu schweren und zu leichten).
- Die Ergebnisse eines Tests werden streuen, wenn alle Aufgaben ein mittleres Trennschärfeniveau aufweisen.

Das zweite Kriterium, die Trennschärfe, ist noch nicht bekannt, wird aber im nächsten Schritt vorgestellt.

6.4. Trennschärfe

Der Begriff *Trennschärfe* ist zunächst nicht einfach zu verstehen. Im Kern bezeichnet er die Fähigkeit einer Aufgabe, geeignet zum Gesamtergebnis beizutragen (Bortz 1995). Dieser Zusammenhang zwischen Aufgabe und Testergebnis wird mathematisch als Korrelationswert zwischen einer Aufgabe und den restlichen Aufgaben dargestellt. Dabei bezeichnet der Begriff *Korrelation* eine Größe, die den mathematisch-linearen Zusammenhang zwischen zwei Variablen beschreibt.

Jede Korrelation hat einen Wertebereich von –1 bis 1. Dabei gelten folgende Interpretationsregeln:

- Eine Korrelation von –1 bezeichnet einen idealen negativen Zusammenhang, das heißt: Personen, die hohe Werte in Variable 1 haben, haben automatisch niedrige Werte in Variable 2.
- Eine Korrelation von 1 bezeichnet einen idealen positiven Zusammenhang, das heißt: Personen, die hohe Werte in Variable 1 haben, haben automatisch hohe Werte in Variable 2.
- Eine Korrelation von 0 besagt, dass die beiden Variablen überhaupt nicht miteinander zusammenhängen. Aus der Ausprägung von Variable 1 kann nicht auf die Ausprägung von Variable 2 geschlossen werden.

Ideale Zusammenhänge kommen allerdings in der Praxis nur extrem selten vor. Die meisten Korrelationen haben einen mittleren Betrag. Beispielsweise gibt es einen guten mathematischen Zusammenhang zwischen Schulbildung und Einkommen, auch wenn es natürlich einige Fälle gibt, bei denen eine entsprechende Vorhersage scheitert. An dieser Stelle sei kurz angemerkt, dass Korrelation und Kausalität nicht viel miteinander

zu tun haben. Beispielsweise gibt es eine hohe Korrelation zwischen der Anzahl von Storchenpärchen in einer Gemeinde und der Anzahl an Kindern pro Paar in dieser Gemeinde. Daraus zu schließen, dass der Storch die Kinder bringt, ist falsch; es gibt viel mehr einen gemeinsamen Hintergrund, der beide Variablen erklärt: Die Ländlichkeit der Gemeinde.

Für die Korrelation zwischen Aufgabe und Testergebnis stellt sich die Frage der Kausalität jedoch nicht. Hier geht es alleine um die Frage des mathematischen Zusammenhangs, also darum, wie viel die Aufgabe zum eigentlichen Testergebnis beiträgt. Auch hier gibt es einen idealen Korridor, der sich aus folgenden Annahmen herleiten lässt:

- Aufgaben mit einer Trennschärfe von 0 tragen nichts zum Testergebnis bei. Diese Trennschärfe dürfte sich ergeben, wenn Sie in einer Mathematikprobe nach der Lieblingsfarbe der Schüler/-innen fragen.
- Aufgaben mit negativer Trennschärfe wirken dem Testergebnis sogar entgegen: Schlechte Schüler/-innen erhalten viele Punkte, gute weniger. Das kann passieren, wenn ein an sich falscher Denkfehler schnell zur richtigen Lösung führt. Schlechte Schüler/-innen verwenden ihn und lösen die Aufgabe, gute erkennen ihn, kommen dann jedoch nicht weiter.
- Aufgaben mit positiver Trennschärfe tragen zum Testergebnis bei: Gute Schüler/-innen schneiden in der Aufgabe besser ab als schlechte. Hier gilt jedoch:
- Aufgaben mit sehr hoher positiver Trennschärfe führen den gesamten Test ad absurdum. Wenn das Testresultat bereits durch eine einzige Aufgabe ermittelt werden kann, wird der restliche Test nicht mehr benötigt.

Daher gilt die Faustregel: Jede Aufgabe sollte eine Trennschärfe im mittleren positiven Bereich, also von 0,3 bis 0,8 aufweisen.

Die Einhaltung der bis hierhin genannten Kriterien für einen guten Test (Verteilung, Schwierigkeit, Streuung, Aufgabentrennschärfe) kann bei der eigentlichen Testerstellung noch nicht geprüft werden. Gerade in den ersten Jahren des Lehrberufs ist es jedoch sinnvoll, bei allen Tests die entsprechenden Kenngrößen zu prüfen und diese als Erfahrungswert für die nächsten Tests produktiv zu verwenden. Auf diese Weise kommt man dem Ziel, gute Tests zu erstellen, deutlich näher. Als Beispiel mag die Episode 1 dienen: Aufgabe 1 scheint zu schwer, Aufgabe 3 nicht trennscharf. Diese Hinweise, den jeweiligen Aufgabenstellungen gegenübergestellt, ergeben gute Reflexionsanlässe und Erfahrungswerte für die nächsten Tests.

Die nächsten drei Gütekriterien (Objektivität, Reliabilität und Validität) sind die bekannteren der Testgütekriterien. Sie spielen nicht nur bei der Durchführung und Auswertung der Tests eine große Rolle, sondern bereits bei der Testerstellung und müssen entsprechend von Anfang an berücksichtigt werden (Arnold 2002; Heller/Hany 2002; Weinert 2002).

6.5. Objektivität

Unter dem Begriff Objektivität versteht man die Vorgabe, dass ein Test unter Rahmenbedingungen stattfinden muss, die die Leistung der Testteilnehmer nicht beeinflussen (Bortz 1995; Arnold 2002). Das bedeutet, dass ein Test dann objektiv ist, wenn die Testergebnisse möglichst unabhängig von den Bedingungen am Prüfungsort und von der Person sind, die den Test beaufsichtigt und auswertet. Es muss also theoretisch gleichgültig sein, ob die Lehrkraft selbst oder eine Vertretung den Test auswertet.
Dies ergibt eine Reihe von Handlungsempfehlungen für die Durchführung, die Korrektur und die Notenvergabe von Tests in der schulischen Praxis. Man spricht in diesem Zusammenhang von der Durchführungsobjektivität, der Auswertungsobjektivität und der Interpretationsobjektivität: Sowohl die Situation der Prüfung, als auch die Korrektur der Antworten, als auch die Bewertung der Gesamtleistung müssen objektiv durchgeführt werden. Für jeden dieser Schritte solten möglichst viele Handlungen bereits im Vorfeld des Tests festgelegt und standardisiert werden.

Durchführungsobjektivität

Die Sicherung der Durchführungsobjektivität dient dazu, allen Schüler/-innen möglichst identische Testbedingungen zu ermöglichen. Dies wird in der Regel dadurch erreicht, dass man die Bearbeitungszeit vereinheitlicht, die Aufgabenstellungen für alle verständlich gestaltet oder den Gebrauch von Hilfsmitteln für alle Schüler/-innen zulässt. Diese Regeln sind Standard im deutschen Schulwesen, keine Lehrkraft käme auf die Idee, Lernende während eines Tests unterschiedlich zu behandeln. Es gibt aber einige Aspekte, die im Rahmen der Durchführungsobjektivität oftmals übersehen werden. Diese sind:

- **Äußere Faktoren.** Ist die Prüfung in der ersten Stunde oder in der letzten? Ist es im Klassenzimmer für die Schüler/-innen, die am Fenster sitzen, zu heiß?
- **Hilfestellungen.** Stellt jemand im Verlauf einer schriftlichen Prüfung eine Frage, so muss die Antwort auf diese Frage entweder nicht gegeben werden oder alle Schüler/-innen gemeinsam informiert werden.
- **Unerlaubte Hilfsmittel.** Unternimmt man als Aufsichtsperson während einer Prüfung nicht den ernsthaften Versuch, unerlaubte Hilfsmittel zu entdecken, schafft man damit ebenfalls ungleiche Testbedingungen: Die Schüler/-innen mit unerlaubten Hilfsmittel sind im Vorteil, auch wenn es die Lehrerin der Episode 2 anders sieht.
- **Positionswechsel.** Auch wenn es aus eigener Erfahrung enervierend ist: Bleibt die Lehrkraft stets am Pult sitzen, fühlen sich nur die Schüler/-innen in den ersten Reihen streng beobachtet. Häufige Positionswechsel hingegen verteilen das Gefühl des Beobachtetwerdens besser über alle Beteiligten.

Durchführungsobjektivität ist natürlich nur näherungsweise zu erreichen. Verschiedene Aspekte wie Sitzposition, Reaktionen auf Äußerungen der Lehrkraft während der Prüfung oder die Funktion des Schreibgerätes sind nicht kontrollierbar und damit auch nicht objektivierbar. Das Ziel muss es daher sein, die Durchführung einer Prüfung so weit wie möglich zu objektivieren, die Bedingungen im Vorfeld standardisiert festzulegen, bekanntzugeben und die Prüfung bei einem weitgehend störungsfreien Verlauf das nächste Mal auf ähnliche Weise abzuhalten. Kiel (2001) spricht in diesem Zusammenhang von kontrollierter Subjektivität als Ziel einer Prüfungsdurchführung.

Ob die Durchführungsobjektivität eingehalten wird, kann letztendlich erst nach der Durchführung der Prüfung konkret untersucht werden. Jedoch kann und sollte gerade eine unerfahrene Lehrkraft bei den ersten Tests im Vorfeld das Gespräch mit den Kolleg/-innen suchen. Dabei sollte man sich nicht nur fremde Erfahrungen abholen, sondern vielmehr die eigenen Ideen schildern und diese kritisch evaluieren lassen.

Auswertungsobjektivität

Unter Auswertungsobjektivität versteht man, dass alle Tests nach gleichen Standards und unter gleichen Bedingungen ausgewertet werden. Das spricht im Kern die bereits in Kapitel 2 ausgeführten Fehler bei der Beschreibung und Interpretation von Prüfungshandlungen an. Hier geht es um die Frage, wie die Bewertungsgrundlage der unterschiedlichen Prüfungsleistungen hergeleitet wird. Es gelten im Wesentlichen die bereits beschriebenen Regeln zur Sorgfalt und Genauigkeit bei der Auswertung, hier speziell:

- Das Erstellen von Kriterien, wie die einzelnen Beobachtungen zu interpretieren sind.
- Die genaue und stets gleichbleibende Anwendung des Kriterienkatalogs.
- Die Konzentration auf den Inhalt und nicht auf die Rahmenbedingungen.
- Kenntnis und konstante Berücksichtigung der verschiedenen Fehlerquellen (Kapitel 2).

Hierzu ein Beispiel: Um Reihungseffekte oder Voreinschätzungen zu minimieren, könnte man einen schriftlichen Test so gestalten, dass jede Aufgabe auf einem eigenen Blatt bearbeitet wird. Die erste Seite bildet ein Deckblatt, auf dem der Prüfling lediglich seinen Namen einträgt. Bei der Korrektur schlägt man die erste Seite um, mischt die schriftlichen Arbeiten, korrigiert jeweils die erste Aufgabe nach dem Kriterienkatalog und mischt den Stapel vor der Korrektur der nächsten Aufgaben erneut. Erst am Schluss addiert man die Punkte.

Eine weitere Möglichkeit zur Steigerung der Auswertungsobjektivität, die der an dieser Stelle erwähnt werden muss, ist die Maximierung der Fragenkontingenz. Darunter versteht man das Verwenden von

Fragestellungen, die lediglich durch das Hinschreiben eines Wortes oder durch das Ankreuzen eines Kästchens zu lösen sind. Bei solchen Antworten ist eine objektive Auswertung sehr einfach umzusetzen. Dieses Verfahren muss jedoch kritisch gesehen werden, da es eine Menge unerwünschter Konsequenzen mit sich bringt; insbesondere können mit diesem Verfahren nur niedrigwertige kognitive Leistungen, meist nur Behaltensleistungen, getestet werde (siehe Kapitel 5).

Interpretationsobjektivität

Die Sicherung der Interpretationsobjektivität betrifft in erster Linie den Vorgang des Überführens der festgehaltenen Erkenntnisse in eine begründete Bewertung: Gleiche Ergebnisse müssen zur gleichen Note führen. Damit ist sie, nimmt man das BIB-Modell von Kiel (2001) als Grundlage, im Bereich der weitergehenden Interpretation und der Bewertung verortet. Bei der Interpretationsobjektivität ist eine hohe Objektivität gewährleistet, wenn gleichwertige Wissensstände zu gleichen Beurteilungen führen, wenn also gleiche Punktestände unabhängig von Sympathieeffekten oder anderen Tendenzen zu gleichen Noten führen.

Ein über einen reinen Kriterienkatalog hinausgehender Zuordnungsschlüssel von Leistungen zu Bewertungen ist dabei hilfreich, solche Fehler zu vermeiden. Es ist aber nicht möglich, einen solchen für alle Prüfungsgelegenheiten zu erstellen. Dennoch ist es empfehlenswert, neben den Kriterien auch eventuelle Abstufungen festzuhalten, sofern die untersuchten Kompetenzen dies zulassen. Ist dies nicht der Fall, empfiehlt es sich, oft vorkommende Fehlerquellen, aber auch Aspekte, die einer Prüfungsleistung zusätzlichen Glanz verleihen, festzuhalten und diese als Grundlage der eigenen Bewertung zu nutzen. Das Beispiel eines Kriterienkatalogs in Kapitel 7 nimmt diese Empfehlung auf. Ebenso lohnt sich eine gründliche Evaluation des eigenen Bewertungsverfahrens im Kollegenkreis – dies natürlich am günstigsten im Vorfeld einer Prüfung.

Grundsätzlich ist im Kontext der Sicherung der Objektivität festzuhalten: Je höherwertiger oder unterschiedlicher die zu prüfenden kognitiven Leistungen sind, desto schwieriger ist es, Kriterien und Bewertungsgrundlagen standardisiert aufzustellen und ausschließlich diese als Grundlage der eigenen Bewertung zu verwenden (Bügelmann et al. 2006). Bei den eigenen Objektivitätsbemühungen muss es jedoch das Ziel sein, möglichst ohne eigene Einschätzung und Befindlichkeiten zu argumentieren, die Ergebnisse vorurteilsfrei und nachvollziehbar abzuleiten und die Begründung der Note sach- und handlungsorientiert zu verfassen.

6.6. Reliabilität

Unter der Zuverlässigkeit oder Reliabilität einer Messung versteht man den Grad der Sicherheit oder Genauigkeit, mit dem ein bestimmtes Merkmal gemessen werden kann (Bortz 1995; Hartig/Klieme 2007). Diese Definition schließt an Aussagen der klassischen Testtheorie an und beruht auf dem Versuch, insbesondere den systematischen Fehler des eigentlichen Messinstruments, also zum Beispiel einer schriftlichen Aufgabenstellung, möglichst klein zu halten. Ein Messinstrument mit einer hohen Reliabilität sollte also, wenn man den gleichen Test mehrmals ausführt, immer das gleiche Ergebnis liefern. Die Empfehlungen der empirischen Sozialforschung zur Sicherstellung der Reliabilität sind im Schulkontext nur zu einem kleinen Teil nutzbar. Dort werden beispielsweise Messwiederholungen gefordert, sodass der Grad der Reliabilität eines Messinstruments, beispielsweise eines Fragebogens, an Kennzahlen festgemacht werden kann, die auf einem Vergleich der beiden Messreihen basieren. Zwar ist es grundsätzlich auch in einer schriftlichen Prüfung möglich, entsprechende Kennzahlen zu ermitteln (beispielsweise den Cronbachs Alpha oder die Konstruktreliabilität), allerdings ist die Problemstellung der Reliabilitätssicherung bei schulischen Tests nur am Rande eine mathematische. Zudem kann beispielsweise der Cronbachs Alpha künstlich erhöht werden, indem die Fragen des Tests so formuliert werden, dass sie sich nur oberflächlich unterscheiden. Ein hoher mathematischer Reliabilitätswert ist schön, darf jedoch nicht um jeden Preis zu erreichen versucht

werden, da sonst die Brauchbarkeit des Instrumentes für den schulischen Einsatz leidet.

In der Schule geht es eher um die Präzision eines Messinstruments, also den Grad der Genauigkeit, mit dem das Ergebnis gemessen werden kann. Diese Genauigkeit wird vom Instrument unterstützt, wenn es bei jedem Prüfling möglichst die gleichen Handlungen anstößt, wenn also jeder zu Prüfende beim Lesen der Fragen sehr genau versteht, wie die geforderten Handlungen bzw. wie das geforderte Ergebnis aussehen sollte. Reliabilität bei einer schulischen Prüfung bedeutet also eine klare und konkrete Aufgabenformulierung, die keine Fragen oder Unklarheiten offen lässt.

Damit ergibt sich als Konsequenz, dass es bei der Reliabilität eines für schulische Prüfungen verwendeten Tests darauf ankommt, dass Fragen präzise und eindeutig gestellt werden, sodass in der Frage bereits die Erwartung integriert ist. Dies ist auch für schlecht definierte, hohe kognitive Leistungen erfordernde Aufgabenstellungen möglich, wie es das folgende Beispiel zeigt:

- Eine reliable Frage könnte sein: Sind, wie Mephisto Faust vorwirft, Liebesschwüre potenzielle Lügen? Diskutieren Sie diesen Vorwurf unter Bezugnahme auf die Szene „Straße“ (3025–3072) in Goethes Faust.
- Weniger reliabel ist folgende Frage: Diskutieren Sie, ob die Hauptfigur in Goethes Faust fremd- oder selbstbestimmt handelt. (Welche Figur ist gemeint?)

Zur Sicherung der Reliabilität kann an dieser Stelle wiederum die bekannte Empfehlung gegeben werden, die eigenen Tests im Kollegenkreis evaluieren zu lassen und insbesondere zu prüfen, ob die Intention der Fragestellungen eindeutig ist.

6.7. Validität

Die Validität gibt als wichtigstes Gütekriterium eines schulischen Tests an, ob ein Test auch das misst, was er messen soll (Bortz 1995; Moosburger/ Kelava 2007). Das zur Illustration dieser Definition oft zitierte Beispiel einer Textaufgabe in der Mathematik, die so kompliziert gestellt ist, dass viele Schüler/-innen den Sinn der Aufgabe nicht begreifen, ist allerdings eher ein Beispiel für schlechte Reliabilität. Die Validität im schulischen Kontext bezieht sich eher auf den zu prüfenden Stoff.

In der empirischen Sozialforschung (Bortz/Döring 1995) werden drei unterschiedliche Formen der Validität unterschieden:

- **Inhaltsvalidität.** Die Inhaltsvalidität wird angenommen, wenn ein Verfahren zur Messung eines bestimmten Konstrukts bzw. Merkmals alle Aspekte dieses Konstrukts ausreichend ausschöpft. So sollte z. B. ein Test zur Rechenleistung zumindest alle Grundrechenarten enthalten und sich nicht nur auf Addition und Subtraktion beschränken.
- **Kriteriumsvalidität.** Von Kriteriumsvalidität wird gesprochen, wenn als Kriterium ein anderer, als valide anerkannter Test herangezogen wird und die Messdaten der beiden Tests hoch miteinander korrelieren.
- **Konstruktvalidität.** Von einer hohen Konstruktvalidität spricht man, wenn davon ausgegangen werden kann, dass der Bedeutungsumfang des Konstruktes präzise und nachvollziehbar abgebildet ist. Das ist der Fall, wenn die Messdaten von Tests, die dasselbe Konstrukt abbilden, hoch miteinander korrelieren (konvergente Validität) und die Messdaten von Tests, die verschiedene Konstrukte abbilden, nur gering miteinander korrelieren (diskriminante Validität).

Eine mathematische Überprüfung der Validität, wie sie in der empirischen Sozialforschung durchaus üblich und geboten ist, ist bei schulischen Tests jedoch nicht erforderlich. Hier geht es vielmehr um die Sicherstellung der Inhaltsvalidität, in diesem Zusammenhang oft auch als *curriculare Validität* bezeichnet. Das heißt: Sie müssen das prüfen, was laut Lehrplan geprüft werden soll und was Sie im Unterricht ausreichend vorbereitet haben. Darüber hinaus sollten Sie darauf achten, dass diese Inhalte in den Fragen angemessen gewichtet sind und nicht nur ein kleiner, willkürlich ausgewählter Teil der Inhalte abgeprüft wird – sollten Sie etwa in Geographie die Rolle der Industrie in Deutschland an Hand von fünf Beispielen erklärt haben und schreiben eine Prüfung, die sich nur mit der Rolle der Automobilindustrie beschäftigt, können Sie die Kenntnis des Stoffes nicht valide überprüfen.

Zur Sicherung der curricularen Validität empfiehlt sich eine Prüfung der Testitems mit mehreren Experten, z. B. Kolleginnen und Kollegen aus dem gleichen Fach. Zusätzlich sollte man stets

- anforderungsbezogen vorgehen,
- vorhandene wissenschaftliche Erkenntnisse nutzen,
- Objektivität und Reliabilität sicherstellen und
- die Tests evaluieren und verbessern.

Der letzte Punkt heißt konkret: Evaluieren Sie, was laut Lehrplan drankommen muss und überprüfen Sie die curriculare Validität Ihrer Tests durch Gespräche mit Kollegen und gegebenenfalls mit dem Fachbereichsleiter.

6.8. Weitere Gütekriterien

In der Literatur (Arnold 2002; Bortz/Döring 1995; Hartig/Klieme 2007; Moosburger/Kelava 2007) werden zu den genannten Gütekriterien weitere Aspekte eines guten Tests genannt. Die bekanntesten dieser Kriterien werden im Folgenden kurz vorgestellt:

- **Ökonomie.** Ein Test ist dann ökonomisch, wenn er mit vertretbarem Aufwand für alle Beteiligten durchgeführt werden kann, das heißt, wenn die Durchführung, die Auswertung und die Interpretation der schriftlichen Arbeit in angemessener Zeit und mit überschaubarer Anstrengung durchgeführt werden kann.
- **Zumutbarkeit.** Ein schulischer Test wird als zumutbar bezeichnet, wenn er den zu Beurteilenden nicht über alle Maßen hinaus physisch oder psychisch beansprucht. Überspitzt formuliert heißt das: Das aus amerikanischen Strafprozessen bekannte Kreuzverhör, das oftmals über Stunden geht und viele unangenehme Fragen enthält, die einen Zeugen oder Angeklagten psychisch sehr belasten, ist keine geeignete Blaupause für schulische Prüfungen. Das gleiche gilt für Prüfungen der körperlichen Tüchtigkeit eines Jugendlichen, wenn diese eher an die Aufnahmeprüfung für Kampftaucher erinnern.
- **Fairness.** Ein guter Test muss selbstverständlich auch fair sein. Es muss also für jeden Prüfling eine reelle Chance geben, den Test zu bestehen. Prüfungen, die von vornherein darauf ausgelegt sind, eine bestimmte Quote an Kandidaten „auszusieben" (wie es im universitären Kontext fälschlicherweise gerne als das eigentliche Ziel einer Prüfung kolportiert wird), sind demnach nicht fair.

6.9. Die Macht des Zufalls

Manchmal tritt jedoch der Fall ein, dass trotz gründlicher Vorbereitung und trotz genauer Beachtung der Gütekriterien die Ergebnisse eines Tests nicht den Erwartungen entsprechen. Nimmt man als Beispiel die Episode 3 dieses Kapitels, so ist es verständlich, dass der Lehrer nach Gründen sucht, die das vorliegende Ergebnis erklären. Die Wahrscheinlichkeit, dass er diese Gründe findet, ist gering, denn

- Herr Schütze ist ein erfahrener Lehrer.
- Eine Veränderung seiner Handlungen ist nicht festzustellen.
- Das Ereignis ist so noch nie aufgetreten.

Ein ungewöhnliches Testergebnis bedeutet noch lange nicht, dass die Schulaufgabe deswegen nicht angemessen ist. Die Schüler/-innen in der betreffenden Englischschulaufgabe könnten zum Beispiel generell schlechter sein als die in den vorherigen Klassen, was sich relativ schnell durch zurate ziehen der letztjährigen Notenspiegel klären lässt. Zudem könnten äußere Umstände eine geringere Lern- oder Konzentrationsbereitschaft bedingen (z. B. wenn die Klasse in derselben Woche schon mehrere Leistungsnachweise hinter sich hat oder sich eine lärmende Baustelle direkt am Grundstück nebenan befindet). Letztendlich gibt es eine Reihe von Gründen außerhalb der Gestaltung des Tests, die das Ergebnis erklären können; meistens bleiben diese Gründe nicht erkennbar. So lange es sich nicht um systematische Gründe handelt, die man meist mit Hilfe einer Reflexion dieses und der folgenden Tests erkennt, bleiben diese

Gründe unerkannt, treten im Folgenden aber auch nicht mehr auf – es handelt sich also um zufällige, unsystematische Fehler. Möglicherweise ist es, wie Herr Schütze es im ersten Satz des Falles beschreibt, einfach mal passiert. Diese Begründung ist wahrscheinlich, denn es werden jede Woche so viele Schulaufgaben in unzähligen Schulen geschrieben, dass ungewöhnliche Ereignisse irgendwann auftreten.

Infolgedessen sollte Herr Schütze dem Schulleiter die Notenverteilung unverfälscht vorlegen. Denn ein Gespräch mit dem Direktor ist, anders als im Szenario dargestellt, keinesfalls verwerflich. In dem Gespräch kann – wahrscheinlich sehr schnell und positiv – die Angemessenheit von Test und Vorbereitung überprüft werden. In der Praxis läuft die Prüfung auf Angemessenheit eines Tests ohnehin nur selten anhand der Kontrolle der Testgütekriterien ab. Eine erfahrene Schulleitung wird sich vielmehr bemühen, anhand von Hefteinträgen einiger Schüler/-innen und in Beratung mit Kollegen desselben Faches die Angemessenheit des Tests zu bewerten. Danach wird man zum Schluss kommen, die Klasse weiter zu beobachten, um festzustellen, ob dieses Ereignis singulär oder systematisch ist. In letzterem Fall ist es dann tatsächlich die Aufgabe des Lehrers, geeignet zu reagieren, ansonsten ist es „halt einfach mal passiert".

Allgemein sollte man sich nicht verrückt machen, wenn ein solcher Fall auftritt. Es wäre falsch, zu Mitteln zu greifen, die eine bewährte pädagogische Arbeit aufgrund eines Einzelfalls in Frage stellen.

6.10. Bilanzierung des Gelernten

In diesem Kapitel werden folgende Lernimpulse gesetzt:

- Ein Testergebnis setzt sich immer aus dem wahren Wert der Merkmalsausprägung und einem Messfehler zusammen. Ein guter Test weiß um den Messfehler und versucht ihn zu minimieren.
- Es gibt zwei Arten von Messfehlern: Zufällige Messfehler sind nicht vorhersehbar und kaum verhinderbar. Sie gleichen sich mit der Zeit jedoch aus. Systematische Messfehler hingegen bleiben bestehen, können aber auch erkannt und eliminiert werden.
- Ein guter Test ermittelt verteilte Ergebnisse und lässt so Qualitätsaussagen zu. Er muss aber nicht normalverteilt sein.
- Ein Test darf weder zu schwer noch zu leicht sein, darf aber auch nicht ausschließlich mittelschwere Aufgaben enthalten, da er sonst nicht ausreichend streut. Um zu streuen, muss er verschieden schwere, trennscharfe Aufgaben enthalten.
- Ein Test muss in Durchführung, Auswertung und Interpretation objektiv sein. Das bedeutet, dass die Bedingungen für alle zu testenden Schüler/-innen in all diesen Punkten möglichst gleich sein müssen.
- Ein Test muss reliabel sein. In der Messtheorie bedeutet das, dass mehrere Durchführungen des Tests zum gleichen Ergebnis führen. In der schulischen Praxis ist ein Test dagegen reliabel, wenn seine Fragen klar verständlich sind, sodass jeder Prüfling exakt weiß, was von ihm erwartet wird.
- Ein Test muss auch valide sein. Unter curricularer Validität versteht man, dass ein Test auch wirklich das misst, was ein Lernender laut Lehrplan können soll und was Sie ihn unterrichtet haben. Daneben muss dieser Stoff angemessen vertreten werden, sodass nicht nur ein winziger, willkürlicher Ausschnitt daraus geprüft wird.
- Es gibt noch mehrere weitere, untergeordnete Gütekriterien, und zu guter Letzt gibt es immer noch den Zufall. Er lässt sich auch mit der besten Planung nicht ausschalten, aber wesentlich ist, dass man bei Beachtung der Testgütekriterien auch bei einem sehr ungewöhnlichen Ergebnis nachweisen kann, dass man seinen Pflichten als professionelle Lehrkraft nachgekommen ist und einen angemessenen, objektiven, reliablen und validen Test vorgelegt hat.

In diesem Kapitel wurden die meisten Aspekte der vorgestellten Episoden bereits ausführlich behandelt. So wissen Sie jetzt, dass Ergebnisse wie das von Herrn Schütze immer wieder einmal auftreten können, und er deswegen keinesfalls nachträglich den Notenschlüssel abändern sollte (vgl. auch die rechtliche Lage laut Kapitel 4). Das Verhalten der Kollegin aus Episode 2 ist ebenfalls nicht nachahmenswert. Abgesehen davon, dass es Vorschriften gegen Unterschleif gibt, wird damit die Güte eines Tests keinesfalls erhöht.
Das Testergebnis aus Episode 1 können Sie jetzt ebenfalls einordnen: Die erste Aufgabe war mit einer

Schwierigkeit von 0,2 gerade noch im Rahmen des akzeptablen, aber schon fast zu schwer; die Aufgabe drei war nicht trennscharf, und insgesamt hatte der Test wohl eine gewissen Tendenz zur Mitte, da weder sehr gute noch sehr schlechte Arbeiten vorkamen. Das sind keine optimalen Ergebnisse; wenn Sie daraus lernen, werden Sie das nächste Mal ihre Aufgaben sorgfältiger vorbereiten: Überprüfen Sie, warum die Aufgabe drei nur von schlechten Schülern gelöst wurde, und versuchen Sie, den Schwierigkeitsgrad Ihrer Aufgaben besser zu durchmischen.

Auch für das Gespräch mit dem Rektor aus Episode 2 des ersten Kapitels sind Sie jetzt gerüstet, solange Sie sich bei der Vorbereitung des Tests professionell verhalten haben: Dass die Notenverteilung unnormal ist, nur weil es keine Dreier gab, ist sehr unwahrscheinlich, und ob die Klausur zu schwer war, können Sie mit Hilfe der vorgestellten Instrumente bestimmen. Dass Sie etwas abgefragt haben, was Sie nicht durchgenommen haben, widerlegen Sie am besten, indem Sie dem Rektor Hefteinträge der Schüler zu dem Stoff zeigen. Dass Sie nur bewerten, was Ihnen gefällt und dass die Fragen nicht klar waren, können Sie widerlegen, wenn Sie transparent und professionell gehandelt haben, wenn sie also beispielsweise getreu dem BIB-Schema Beobachtungen, Interpretationen und Bewertungen konsequent getrennt haben und bei der Erstellung ihrer Fragen Vygotskys Wissenszonen berücksichtigt haben.

Wie Sie Fragen noch besser transparent herleiten und objektiv bewerten können, zeigt Ihnen das folgende Kapitel. Es nimmt die Empfehlungen und Konsequenzen der Testgütekriterien auf und präsentiert eine Handlungsempfehlung für die Korrektur schriftlicher Tests in einem schlecht definierten Themengebiet: die Erörterung.

7. Von der Leistungsbeobachtung zur Note

Bis hierhin wurden in diesem Buch vor allem Regeln und Rahmenbedingungen definiert, die für die Leistungsbeurteilung an Schulen wichtig sind und den damit verbundenen Handlungen einen theoriebasierten Rahmen geben. Ab diesem Kapitel wird die Umsetzung in die Praxis thematisiert. Hierzu werden die eingeführten Richtlinien aufgenommen, konkretisiert und operationalisiert. In diesem Kapitel geht es zunächst um schriftliche Arbeiten in der üblichen Form (z. B. Schulaufgaben), im nächsten Kapitel um die Bewertung von Leistungen in erweiterten Lehr-Lernarrangements (z. B. offener Unterricht).
Eingeleitet werden diese Überlegungen wie immer von drei Episoden:

Episode 1: Gütekriterien in der Praxis
Der Referendar Pösl wendet sich hilfesuchend an seine Seminarlehrerin, denn trotz seiner bisherigen praktischen Tätigkeiten und Erfahrung bereitet ihm die Korrektur von Deutschaufsätzen Probleme. Es fällt ihm besonders schwer, beim Korrigieren die nötige Objektivität walten zu lassen. Daher bittet er sie um Hinweise, wie er bei zukünftigen Deutschaufsatzkorrekturen die nötige Objektivität gewährleisten kann.

Episode 2: Eine andere Art zu korrigieren
Eine Schülerin der elften Klasse schreibt im Fach Französisch zwar sehr lange Aufsätze in den schriftlichen Arbeiten und zeigt sich auch bei der Mitarbeit engagiert, die Ergebnisse ihrer Prüfungen sind jedoch meistens nur zufriedenstellend. Ein Referendar möchte dieser Schülerin helfen und wählt eine etwas andere Art der Aufsatzkorrektur. Er korrigiert denselben Aufsatz drei Mal: Einmal nur nach Rechtschreibung, einmal nach Grammatik und einmal nach Logik. Seine Annahme, dies helfe der Schülerin mehr „als alles rot anzustreichen" wird jedoch enttäuscht, da die Schülerin diese Art der Korrektur anders auffasst. Sie fühlt sich angegriffen und extra streng bewertet. Im Gespräch mit der Schülerin erfährt der Referendar, dass ihr Anliegen primär ein gutes Notenbild ist, weniger der Wissenserwerb. Nun stellt sich der Referendar zwangsläufig die Frage, ob sein Handeln gerechtfertigt bzw. korrekt war.

Episode 3: Korrektes Ausfragen
Das Abfragen einer Schülerin oder eines Schülers zu Beginn der Stunde ist ein übliches Verfahren im Mathematikunterricht an Ihrer Schule. Sie fragen Ihren erfahrenen Kollegen um Rat. Er empfiehlt Ihnen folgendes Vorgehen: Sie schreiben sich in der Vorbereitung der Stunde einfach ein paar Fragen auf, von denen sie glauben, dass sie geeignet sind. Diese Fragen stellen sie. Ihr Bauchgefühl sagt Ihnen aber, dass durch dieses Verfahren die Bewertung nicht zweifelsfrei hergeleitet werden kann. Wem glauben Sie?

Schrader und Helmke (2001) definieren die eigentliche Problemstellung bei der Leistungsbeurteilung sehr anschaulich: Einerseits haben Lehrkräfte in der Regel eine sehr gute Fähigkeit darin, Leistungsunterschiede innerhalb einer Gruppe zu erkennen. Es fällt ihnen aber meistens schwer, Leistungsunterschiede klassenübergreifend einzuschätzen. Diese Einschätzung ist jedoch notwendig, da damit Selektions- und Allokationsentscheidungen begründet werden (zum Beispiel die Zulassung zu einem Universitätsstudium). Im amerikanischen Schulsystem versucht man, dieser Herausforderung durch standardisierte, konvergente Testverfahren am Ende eines Schuljahres zu begegnen. Diese Vorgehensweise bringt jedoch ebenfalls eine Reihe unerwünschter Konsequenzen mit sich (vgl. Kapitel 5). Vieles spricht dafür, zur Lösung die Vorzüge der beiden Verfahren zu kombinieren und die Urteile der Lehrkräfte durch wissenschaftlich abgesicherte und standardisierbare Instrumente zur übergreifenden Leistungsdiagnostik zu ergänzen. Diese Instrumente und Anforderungen wurden in den letzten Kapiteln eingeführt. Zusammengefasst geht es darum,

- die eigenen Bewertungshandlungen transparent zu gestalten,
- Kriterien für die Bewertung von Leistungen der Schüler/-innen aufzustellen und
- unter Beachtung der Gütekriterien Tests zu erstellen und auszuwerten.

Vieles spricht dafür, dass die Wahrscheinlichkeit für eine gelungene Umsetzung dieser Anforderungen mit

der Qualität der Verfahrensbeschreibung, der Einbettung der diagnostischen Standards und der Konzentration auf die Einhaltung der Regeln steigt. Was dieser Satz für die Operationalisierung von Leistungsmessung für die schulische Praxis bedeutet, soll im Folgenden geklärt werden. Grundsätzlich gilt (im Sinne des BIB-Modells von Kiel 2001):

- Transparenz wird gefördert, wenn eine *Beschreibung* der Lernziele und der damit verbundenen Handlungen erfolgt, ergänzt um Kriterien, mit denen die Leistungen bewertet werden.
- Die Einhaltung der Gütekriterien wird gefördert, wenn diese Beschreibungen in Aufgabenstellungen umgesetzt werden und die *Interpretation* der dadurch angestoßenen Leistungshandlungen vor dem Hintergrund der Leistungsanforderungen und unter Bewusstwerdung der möglichen Fehlerquellen erfolgt.
- Die sichere *Bewertung* von Leistungshandlungen wird gefördert, wenn die Zuordnung der Leistungsinterpretationen zu einem Wert nach einem inhaltlich standardisierten Bewertungsschema erfolgt.

7.1. Beschreiben von Leistungsanforderungen

Oelkers (2002) definiert es als wesentliches Ziel der Transparenz, den Schüler/-innen die Verknüpfung zwischen dem Lehrplan, also dem behandelten Stoff, und den Prüfungsanforderungen zu ermöglichen, ihnen also Hinweise darauf zu geben, welche Leistungen abgeprüft werden, sodass das Lernpensum kalkuliert werden kann.

Das bedeutet nicht, es den Lernenden besonders einfach zu machen, indem sehr genau definiert wird, was abgefragt wird. Transparenz heißt vielmehr, die in einer Prüfung erwarteten kognitiven Leistungen so genau zu definieren, dass die Lernenden diese Leistungen geeignet vorbereiten können. Der hierfür notwendige zusätzliche Aufwand hält sich in Grenzen, wenn der Unterricht stets zielbezogen auf zu erwerbende Kompetenzen ausgerichtet und vorbereitet wird.

Oelkers (ebd.) vereint den gesetzlichen und den testtheoretischen Rahmen der schulischen Leistungsbeurteilung und fordert präzise und nachvollziehbare Bewertungshandlungen, so dass daraus ein echter Mehrwert entsteht. Er wendet sich damit auch gegen die These von Ingenkamp (1971), „dass die Zensuren keine Vergleichsfunktionen bei schulexternen Adressaten erfüllen können und dass damit unser gesamtes schulisches Berechtigungswesen auf einer Fiktion beruht“ (S. 192); sein Verständnis von Bewertung lautet vielmehr:

„Versteht man unter ‚Noten‘ die Beschreibung von Leistungen im Blick auf Aufgaben während einer bestimmten Periode der Beurteilung, dann stellen sich eigentlich nur drei wirkliche Probleme:

- Wie vergleichbar sind die Aufgaben?
- Wie transparent sind die Beurteilungen?
- Wie berechnet sich die Lernzeit und so die Chance, die Aufgabe zu lösen?
- […]

Das gelingt am besten durch klar formulierte Standards oder Bezugsnormen, und zwar sowohl der Inhalte als auch der Noten selbst“ (S. 12).

Bewertung in der Schule verlangt damit möglichst eindeutige Systeme zur Beschreibung und Beurteilung von Kriterien, nach denen eine Leistung eingeschätzt wird. Es darf dabei nur um den Vergleich der individuellen Leistung mit einen inhaltlich standardisierten Bezugsmaßstab gehen, der sich auf anerkannte Fachnormen beziehen kann (Oelkers ebd.).

Um die Leistungsanforderungen beschreiben zu können und dadurch transparent zu machen, sind zwei Schritte notwendig:

- In einem *ersten Schritt* muss geklärt werden, welche Leistungen die Lernenden am Ende der Lern- bzw. Vorbereitungseinheit beherrschen müssen. Damit sind zunächst noch keine Handlungen verbunden, es wird lediglich definiert, was gewusst oder gekonnt werden muss. Hierbei helfen die so

genannten *Wissenstaxonomien* nach Bloom (Bloom 1972; Krathwol/Bloom/Masia 1975).

– Der *zweite Schritt* nimmt die einzelnen Leistungen auf und überführt diese in Handlungen. In diesem Schritt wird versucht, möglichst genau zu definieren, was die Lernenden zeigen oder darstellen sollen, um dem Bewertenden nachzuweisen, dass sie es können. Hierbei sind die so genannten *Doppelverb-Lernziele* von Gagné (Gagné/Briggs/Wager 1992) nützliche Hilfsmittel.

Wissenstaxonomien zur Definition der Leistungsziele

Unter dem Begriff *Wissenstaxonomie* versteht man qualitativ abgestufte Leistungsbeschreibungen. Implizit benutzen wir diese Taxonomien bereits seit längerem, wir definieren beispielsweise die Behaltensleistung als die Fähigkeit eines Lernenden, sich an Worte, Vokabeln oder Daten zu erinnern oder die Verständnisleistung als die Fähigkeit, eine Definition in eigenen Worten wiederzugeben. Wissenstaxonomien bilden ein Beschreibungsraster. Sie helfen bei der Darstellung der Leistungsanforderungen, indem sie eine Struktur bereitstellen, die es erlaubt, Leistungen zu gruppieren und mit Sammelbegriffen zu beschreiben. Dadurch fällt es wesentlich leichter, Anforderungen und Ziele zu definieren und das, was gewusst oder gekonnt werden muss, verständlich darzustellen.

Das bekannteste Verfahren zur Taxonomierung von Wissen, d. h. zur qualitativen und strukturierten Beschreibung von Leistungen, die nach Lernprozessen beherrscht werden sollen, stammt von Benjamin Bloom (1972). Dieses Raster unterscheidet die Dimensionen

– kognitive Leistungen (Kopf),
– affektive Leistungen (Herz) und
– psychomotorische Leistungen (Hand)

und stuft die erwarteten Leistungen innerhalb dieser Bereiche in jeweils fünf bis sechs Qualitätsstufen ab, die in der folgenden Tabelle dargestellt und beschrieben werden:

Kognitive Leistungen	Affektive Leistungen	Psychomotorische Leistungen
1. Kenntnis	1. Aufmerksamwerden, Beachten	1. Imitation
2. Verständnis	2. Reagieren	2. Manipulation
3. Anwendung	3. Werten	3. Präzision
4. Analyse	4. Strukturierter Aufbau eines Wertsystems	4. Handlungsgliederung
5. Synthese	5. Erfülltsein durch einen Wert oder eine Wertstruktur	5. Naturalisierung
6. Beurteilung		

Tab. 4: Wissenstaxonomien nach Bloom

Kognitive Leistungen beschreiben Fähigkeiten, die auf schriftliche oder sprachliche Darstellung zielen. Diese Leistungen reichen von einfachen Behaltens- und Erinnerungsleistungen über die Anwendung von Wissensbeständen auf neue Problemstellungen bis hin zur Bewertung fremder Lernergebnisse. Mit dieser Dimension werden Fähigkeiten klassifiziert, die auf Wissen, Denken und Problemlösen abzielen.

Affektive Leistungen beschreiben Einstellungen, Interessen und Werthaltungen, die von der Kenntnis eines Wertes über dessen Aneignung und den Aufbau einer Werte- und Interessensstruktur bis hin zur

Identitätsentwicklung gehen. Diese Leistungen gehen über kognitive Leistungen hinaus, sie beziehen alle Lebensbereiche des Lernenden mit ein und sind eher Handlungsdispositionen.
Psychomotorische Leistungen, die in der Schule beispielsweise im Sport- oder im Werkunterricht benötigt werden, beschreiben motorische Fähigkeiten und Fertigkeiten im Hinblick auf die Koordination des eigenen Körpers oder die sichere und erfolgreiche Bedienung von Werkzeugen. Aber auch die Handschrift wird in der Literatur als wichtige psychomotorische Leistung beschrieben.

Wenn es um die Überprüfung schulischer Leistungen geht, werden überwiegend kognitive Lernziele genannt. Aus diesem Grund beschränkt sich dieses Kapitel auf diese Fähigkeiten. Selbstverständlich sind auch die anderen Dimensionen bei der Leistungsmessung in Schulen. So gibt es etwa für das Fach Sport abgestufte Kriterien zur Leistungsbeurteilung, die mehr als nur die reinen Leistungswerte berücksichtigen (Bayerische Landesstelle für den Schulsport 2014). Bloom differenziert die kognitive Dimension in sechs Kategorien, die nach der angenommenen Schwierigkeit der zu erbringenden Leistung geordnet sind. Man nimmt an, dass höherwertige kognitive Leistungen auch einen größeren und langandauernden Lernprozess notwendig machen, da die mit ihnen verbundenen Aufgabenstellungen nicht nur die Nutzbarmachung von Informationen, sondern auch deren Verknüpfung, Manipulation und Modellierung erfordern. Hinzu kommt, dass höherwertige Problemstellungen meist schlechtdefinierte und/oder komplexe Themenbereiche berühren.

- **Kenntnisse** beinhalten im Kontext kognitiver Leistungen in erster Linie das Behalten und Erinnern von Worten, Definitionen und einzelner Fakten, aber auch die Kenntnis von Zusammenhängen, Kategorien oder Kriterien. Behaltensleistungen sind verhältnismäßig einfach zu beschreiben und zu testen, hier geht es darum, dass der zu Bewertende sich an einfache Begrifflichkeiten erinnern und diese wiedergeben kann. Ein Beispiel für die Behaltensleistung ist: „Die Schülerin/der Schüler kennt die Formel zur Berechnung der Stromstärke aus Spannung und Widerstand."
- **Verständnisleistungen** repräsentieren die Fähigkeit, Zusammenhänge und Sachverhalte in einem spezifischen Kontext mit eigenen Worten wiedergeben zu können oder Unterschiede, Abgrenzungen und Gemeinsamkeiten verschiedener Aspekte zu beschreiben. Ein Beispiel für die Verständnisleistung ist: „Die Schülerin/der Schüler versteht den Unterschied zwischen Sozialisation und Erziehung."
- Die **Anwendung** thematisiert die Fähigkeit, Wissensbestände zur Lösung nicht-trivialer Problemstellungen nutzen zu können. Bekannt sind diese Leistungen auch unter dem Begriff Transfer. Zentral sind hierbei die Fähigkeit zur Abstraktion und Kombination verschiedener Theorien und Gesetzmäßigkeiten, um damit höherwertige Aufgaben zu bearbeiten. Ein Beispiel für eine Transferleistung ist: „Die Schülerin/der Schüler kann ein aussagekräftiges Bewerbungsschreiben verfassen."
- **Analysefähigkeiten** ermöglichen es dem Lernenden, komplexe Problemstellungen daraufhin zu untersuchen, was die Ziele der Aufgabenstellung sind, welche Ausgangsbedingungen vorhanden sind und welche Wege zum Ziel führen. Zudem erlauben es elaborierte Analysefähigkeiten, zu entscheiden, welcher der Wege das Ziel mit der größten Wahrscheinlichkeit oder mit dem geringsten Aufwand erreicht. Ein Beispiel für die Analyseleistung ist: „Die Schülerin/der Schüler findet das günstigste Angebot, mit dem die gesteckten Ziele erreicht werden."
- Die **Synthese** beschreibt das erfolgreiche Handeln in multiperspektivisch zu betrachtenden, das heißt in divergenten Problemlagen. Das sind Aufgabenstellungen, bei denen das Ergebnis immer das Produkt eines Diskurses oder eines Abwägungsprozesses ist. Das schließt Planungsfähigkeiten genauso mit ein wie die Kompetenz, Entscheidungen aufgrund einer Analyse der Vor- und Nachteile bzw. der möglichen Folgen zu treffen. Ein Beispiel für eine Syntheseleistung ist: „Die Schülerin/der Schüler bezieht begründet Stellung zur Wirksamkeit von Arbeitsmarktmaßnahmen."
- Die **Evaluationsleistung** bezeichnet als anforderungsreichste der von Bloom beschriebenen kognitiven Leistungen die Bewertung fremder Anwendungen, Analysen oder Synthesen mit Hilfe von qualitativen Aussagen zu deren Güte oder Brauchbarkeit anhand von gesetzten oder selbst gewählten Kriterien. Ein Beispiel für eine Evaluationsleistung lautet: „Die Schülerin/der Schüler beurteilt die Qualität

von internetbasierten Quellen hinsichtlich ihrer Brauchbarkeit für das Lernthema und stellt seine Gedankengänge nachvollziehbar dar."

Während Behaltens- oder Verständnisleistungen auf relativ einfachem Wege als Bewertungskriterium festgesetzt werden können, fällt es bei komplexen kognitiven Leistungen oftmals schwer, die gewünschten Qualitätsansprüche nachvollziehbar darzustellen. Vollständigkeit ist hier schlicht nicht möglich. Stattdessen sollte man eine möglichst hohe Präzisierung der Erwartungen anstreben, indem man durch die Kombination unterschiedlicher Leistungen einen umfänglichen Erwartungsrahmen aufspannt, der wiederum einen hervorragenden Ausgangspunkt zur Schaffung von Transparenz und zur Erstellung einer Bewertungsgrundlage liefert. Die Basis zur Erstellung der unterschiedlichen Anforderungen bildet dabei immer der Lehrplan bzw. die Kompetenzmodellierung der Bildungsstandards.
Die besondere Schwierigkeit dieses Vorgangs liegt darin, dass die Zahl der Handlungen und Unterscheidungsmerkmale durch die Nicht-Trivialität der meisten schulischen Anforderungen schnell ins Unermessliche wächst (Kiel 2007). Hier gilt es, ein möglichst optimales Maß zwischen Qualität der Darstellung und Handhabbarkeit zu finden.

Operationalisierbare Doppelverb-Lernziele zur Fokussierung der Beobachtung

Gerade in der initialen Berufsphase ist es hilfreich, wenn gleichzeitig zu den Fähigkeiten auch Aktionen festgelegt werden, mit denen die Schüler/-innen ihre Fähigkeiten unter Beweis stellen können. Hierfür eignet sich die Verwendung von so genannten Doppelverb-Lernzielen nach Gagné, Briggs und Wager (1992) in besonderem Maße. Sie stellen eine brauchbare Alternative zu herkömmlichen Lernzielformulierungshilfen dar, da diese die schulische Leistungsbeurteilung nicht hinreichend unterstützen. Eine Definition wie

„Lernziele beschreiben den angestrebten Lerngewinn der Schüler bezogen auf einen bestimmten Inhalt. Sie geben an, welche Ziele der Lehrer mit Hilfe der Unterrichtsthemen erreichen will. Von außen betrachtet bezeichnet man die Kombination von Lehr- und Lernziel als Unterrichtsziel" (Glöckel 1996, S. 97).

ist in der Regel wenig hilfreich bei der Erstellung von Leistungskriterien. Einen ersten Ansatz zur Überführung von Lernzielen in operationalisierbare Kriterien bieten die von Mager (1994) festgelegten Bedingungen für ein gutes Lernziel als

„eine zweckmäßige Zielbeschreibung …, mit der es gelingt, die Unterrichtsabsichten dem Leser mitzuteilen. Ein eindeutig beschriebenes Lernziel ist also eines, mit dem Sie Ihre Absichten erfolgreich mitteilen. Eine gute Zielbeschreibung schließt darüber hinaus eine möglichst große Anzahl möglicher Missdeutungen aus" (Mager 1994, S. 19).

Im Kontext der schulischen Lern- und Leistungsbeurteilung sind Lernziele dann sinnvoll, wenn Sie operationalisierbar sind, das heißt, wenn im Lernziel selbst definiert wird, welche Handlungen bzw. welche beobachtbare Aktion die Lernenden zeigen können, um zu demonstrieren, dass sie das gegebene Lernziel erreicht haben. Diese Präzisierung erfolgt am Besten in Form von so genannten Doppelverb-Lernzielen, wobei man zur Spezifikation des Lernziels bis zu fünf Komponenten festlegt (Gagné/Briggs/Wager 1992).
Die Basis eines jeden Doppelverb-Lernziels bilden Fähigkeit und Aktion, diese werden mit jeweils einem Verb beschrieben. Das heißt:

- Man geht von der bereits formulierten (und nicht direkt beobachtbaren) **Fähigkeit** bzw. kognitiven Leistung aus [Verb 1] und formuliert dazu
- die (beobachtbare) **Aktion,** die der Lernende zeigen soll [Verb 2] sowie optional
- die **Situation,** in der Leistung gezeigt werden soll,
- das **Objekt,** an dem die Leistung gezeigt werden soll,
- **Hilfsmittel,** Beschränkungen oder sonstige spezifische Bedingungen.

Dies ergibt ein Doppelverb-Lernziel, das in jedem Falle ein Kompetenzverb (Fähigkeit) und ein Tätigkeitsverb (Aktion) enthält. Der logische Zusammenhang der beiden so entstehenden Halbsätze wird

durch die Verbindungswörter „indem sie“ hergestellt. Folgendes Beispiel aus dem Biologieunterricht verdeutlicht das System:
Die Schülerinnen und Schüler verstehen *(→ Fähigkeitsverb, nicht direkt beobachtbar)* den Aufbau einer Pflanze, indem sie in Dreier- oder Vierergruppen *(→ Situation)* an vorgelegten echten Pflanzen *(→ Objekt)* Wurzeln, Blatt und Stamm benennen *(→ Aktionsverb, beobachtbar)*. Ein Bestimmungsbuch *(→ Hilfsmittel)* darf hinzugezogen werden.

Generell erstellen Sie ein beliebiges Doppelverb-Lernziel nach folgendem Schema:

- Benennen Sie zuerst die Fähigkeit, die gezeigt werden soll. Orientieren Sie sich dabei am besten an den Bloom’schen Wissenstaxonomien. (Ist die Fähigkeit eine Behaltensleistung? Oder eine Verständnisleistung? Transfer? …)

- Erstellen Sie nun den ersten Halbsatz des Doppelverb-Lernziels, z. B.
  - Die Schülerinnen und Schüler kennen die Vokabeln des Kapitels 1, …
  - Die Schülerinnen und Schüler verstehen den Zusammenhang zwischen Spannung und Stromstärke, …
  - Die Schülerinnen und Schüler wenden die Formel der Flächenberechnung eines Kreises richtig an, …
  - Die Schülerinnen und Schüler analysieren die aktuellen Probleme des Medienkonsums von Jugendlichen, …
  - Die Schülerinnen und Schüler deduzieren ihre individuelle Betroffenheit von der so genannten Klimakatastrophe, …

- Erstellen Sie nun den zweiten Halbsatz mit der Aktion, die gezeigt werden soll, und verbinden Sie diesen durch die Wörter „indem sie“, z. B.
  - … indem sie bei einer Nennung des deutschen Begriffs den entsprechenden englischen Begriff nennen.
  - … indem sie die direkte Proportionalität der beiden Größen beschreiben können.
  - … indem sie die Fläche eines Kreises bei gegebenem Durchmesser korrekt berechnen.
  - … indem sie die Ausgangslage, die Vor- und Nachteile des derzeitigen Medienangebots darstellen.
  - … indem sie aus den allgemeinen Erkenntnissen der Forschung die individuellen Auswirkungen auf ihren Bereich ableiten und diese nachvollziehbar darstellen.

- Fügen Sie gegebenenfalls noch Objekt, Situation und Hilfsmittel hinzu. Wenn keine Hilfsmittel vorkommen, müssen diese natürlich nicht erwähnt werden.
- Der Zusammenhang zwischen Fähigkeitsverben und Aktionsverben lässt sich an folgender Tabelle verdeutlichen.

Kog. Leistung	**Fähigkeit**	**→**	**Aktion**
Kenntnis	Die Lernenden *identifizieren* Pflanzen,	indem sie	vorgelegte Pflanzen mit Namen *benennen.*
Kenntnis	Die Lernenden *unterscheiden* zwischen männlichen und weiblichen Chromosomen,	indem sie	XY-Chromosomen Männern und XX-Chromosomen Frauen *zuordnen.*

Verständnis	Die Lernenden *verstehen* die Bedeutung des § 110 des deutschen Bürgerlichen Gesetzbuchs (BGB),	indem sie	ein Beispiel *nennen,* welches im Einklang mit dem Gesetz steht und ein weiteres Beispiel *nennen,* welches gegen das Gesetz verstößt.
Transfer	Die Lernenden können den Flächeninhalt eines vorgegebenen Dreiecks *bestimmen,*	indem sie	eine dazu geeignete Formel *auswählen* und den Flächeninhalt korrekt *berechnen*.

Tab. 5: Beispiele für den Zusammenhang zwischen kognitiver Leistung, Fähigkeiten und Aktionen

Diese Vorgehensweise ist relativ einfach zu erlernen und anzuwenden, wenn man bei der Formulierung der Aktionen bedenkt, dass die Schülerinnen und Schüler diese auch zeigen müssen. Dies gilt insbesondere dann, wenn die Aktionen komplex und multifaktoriell sind. In diesen Fällen empfiehlt es sich, mehrere Aktionen zu definieren. Zudem ergibt sich der Vorteil, dass die benannten Aktionen eine gute Grundlage für die reliable und valide Gestaltung der eigentlichen Aufgabenstellungen bereithalten. Dennoch sind einige Einschränkungen und Kritikpunkte zu bedenken:

- Es ist nicht immer möglich, exakte Kriterien zu finden, um alle gewünschten Fähigkeiten und Kompetenzen aus den Groblernzielen des Lehrplans zu testen. Insbesondere Aufgaben, die Kreativität oder divergentes Denken erfordern, sind nur sehr schwer exakt zu beschreiben. Hier muss im Extremfall eine kognitive Leistung mit sehr vielen möglichen Aktionen beschrieben werden, was die Ökonomie dieses Verfahrens mindert.
- Die Idee, aus der Verhaltensbeobachtung auf Kompetenzen zu schließen, entstammt ursprünglich dem Behaviorismus. Viele Vorgänge und Prozesse des Lernens sind jedoch durch einfache Beobachtung nicht feststellbar.
- Lernziele sind mit dieser Methode nur dann geeignet, die Leistungsbeurteilung zu unterstützen, wenn die Aktionen direkt beobachtbar sind. Dies ist natürlich eine sehr idealisierte Vorstellung, da Aktionen von unterschiedlichen Personen auch unterschiedlich performt werden. Dies gilt insbesondere für höherwertige kognitive Leistungen.

Gleichwohl bietet die hier vorgestellte Methode einen guten Einstieg in die Umsetzung der schulischen Leistungsbeurteilung.

7.2. Interpretieren von Leistungshandlungen

Wenn Menschen andere Menschen und ihr Handeln beobachten, ist ihre Wahrnehmung durch langfristig erworbene Wahrnehmungsweisen, Motive und Intentionen geprägt und damit selektiv. Der Mensch nimmt subjektiv wahr und wird dies niemals komplett ignorieren können.

„Ein Beobachter scheint hauptsächlich diejenigen Ereignisse bevorzugt aufzunehmen, die er am ehesten in sein persönliches Präferenzsystem einordnen kann; dafür lässt er sich am leichtesten sensibilisieren; bei gegenteiligen Ereignissen findet in der Regel demgegenüber eine Wahrnehmungsabwehr statt. Die Wahrnehmung ist praktisch immer das Ergebnis eines Kompromisses zwischen dem, was der Wahrnehmende erwartet, und dem, was objektiv vorliegt" (Rosemann 1975, S. 180; Schwark 1977, S. 23).

Jedoch schärft im Optimalfall allein das Bewusstsein um diese Subjektivität eine Lehrkraft in solchem Maße, dass sie in der Lage ist, Beschreibungen des Schülerverhaltens von deren Beurteilung, Interpretation und Bewertung zu trennen. Dennoch ist gerade die Interpretation der Beobachtungen von Leistungen eine bedeutende Fehlerquelle, die nur durch bewusstes und wohldurchdachtes Handeln in den Griff zu bekommen ist.

In diesem Zusammenhang muss zunächst die Frage geklärt werden, wie der Mensch zu interpretativen

Schlüssen gelangt. Das kann auf zwei Wegen geschehen: deduktiv und induktiv.

- Die *Induktion* ist der Weg von Einzelbeobachtungen zu einer Schlussfolgerung. Bei dieser Form des Schließens versucht man, durch die Kombination möglichst vieler Einzelbeobachtungen eine Regel oder eine Allgemeingültigkeit zu erschließen. Ein Beispiel:
  - Ich sehe einen weißen Schwan. (Beobachtung)
  - Ich sehe noch einen weißen Schwan. (Beobachtung)
  - Ich sehe noch einen weißen Schwan. (Beobachtung)
  - […]
  - Alle Schwäne sind weiß. (Schlussfolgerung)

  Für das induktive Schließen bei der Bewertung von Leistungen wäre als Beispiel zu nennen:
  - Ich sehe, dass Anna einen Begriff in eigenen Worten definiert. (Beobachtung)
  - Ich sehe, dass Anna eine Anwendung des Begriffs beschreibt. (Beobachtung)
  - Ich schließe daraus, dass Anna das notwendige Wissen verfügt. (Schlussfolgerung)

Dieses Vorgehen ist intuitiv und stützt sich nicht direkt auf Kriterien und zu beobachtende Aktionen, die im Vorfeld festgelegt wurden.

- Die *Deduktion* ist der Weg von der Theorie oder von einem Axiom (vom Allgemeinen oder Bekannten) zum Einzelfall. Dieser Weg, zu Erkenntnissen oder Schlussfolgerungen zu gelangen, nutzt einen Satz, der als wahr angenommen wird und damit die Regel für die Schlussfolgerung bildet. Aufgrund von Fällen oder Beobachtungen werden die Schlussfolgerungen abgeleitet. Ein Beispiel hierzu ist:
  - Alle Menschen sind sterblich. (Regel)
  - Ich bin ein Mensch. (Beobachtung)
  - Ich bin sterblich. (Schlussfolgerung)

  Für das deduktive Schließen bei der Bewertung von Leistungen wäre als Beispiel zu nennen:
  - Eine Schülerin / ein Schüler kann den Flächeninhalt eines vorgegebenen Dreiecks bestimmen, wenn sie die korrekte Formel auswählt, die vorgegebenen Größen einsetzt und das Ergebnis korrekt berechnet. (Regel)
  - Anna nennt die korrekte Formel. (Beobachtung)
  - Anna wählt die entsprechenden Werte zu den Elementen der Formel. (Beobachtung)
  - Anna berechnet ein korrektes Ergebnis. (Beobachtung)
  - Ich schließe daraus, dass Anna den Flächeninhalt eines vorgegebenen Dreiecks bestimmen kann. (Schlussfolgerung)

Dieses Vorgehen lehnt sich eng an die operationalisierten Doppelverb-Lernziele an und benutzt sie als Regel, die durch Beobachtungen zu falsifizieren oder zu stützen ist.

Bei der schulischen Leistungsbeurteilung schließt die Lehrkraft durch Beobachtungen auf Fähigkeiten. Laut Popper (1935/2005) ist das induktive Schließen der falsche Weg, um zu gültigen Erkenntnissen zu gelangen. Auch durch noch so viele Einzelbeobachtungen wird es nicht gelingen, Wahrheiten oder auch nur allgemeingültige Regeln zu erschließen, denn:

- Induktives Schließen nutzt keine Hypothesen oder Axiome als Basis des Erkenntnisgewinns, diese werden vielmehr erst aus – mehr oder weniger – zufälligen Einzelbeobachtungen erschlossen.
- Ob die Einzelbeobachtungen ausreichen, um zu einem sicheren Ergebnis zu gelangen, kann nicht geklärt werden.
- Die Schlussfolgerungen als wahr anzunehmen ist daher schlichtweg nicht möglich.

Hinzu kommt, und hierin liegt ein großes Problem, dass der Mensch die fundamentale Tendenz hat, immer nur denjenigen Beobachtungen besondere Aufmerksamkeit zu schenken, die den nach einigen Einzelbeobachtungen getroffenen Schlussfolgerungen entsprechen. Diese Annahme konnte Wason (1960) durch ein Aufsehen erregendes Experiment gut illustrieren.
Wason lud dazu mehrere Probanden ein, an einem Spiel teilzunehmen. Die Regeln dieses Spiels lauten:

- Die Versuchsleitung nennt ein Zahlentripel, zum Beispiel 2-4-6. Dieses Zahlentripel gehorcht einer Regel, die nur die Versuchsleitung kennt.
- Die Aufgabe der Probanden ist es nun, diese Regel durch Einzelbeobachtungen herauszufinden, sie also durch eine Kombination mehrerer Erkenntnisse induktiv zu schließen.
- Die Probanden nennen nun ihrerseits Zahlentripel, von denen sie glaubten, dass es ihnen hilft, die Regel zu verstehen.
- Die Versuchsleitung antwortet auf jedes genannte Zahlentripel mit ‚Richtig' (= das Zahlentripel gehorcht der Regel) oder ‚Falsch' (= das Zahlentripel gehorcht der Regel nicht).
- Die Probanden wiederholen dies, bis sie glauben, die Regel zu kennen.

Die diesem Spiel zugrundeliegende Regel lautet: Die Zahlen unterscheiden sich. Die Regel ist trivial, allerdings gibt es nur einen Weg, diese Regel zu erschließen, und diese Regel ist deduktiv: Die Probanden müssen sich eine willkürliche Regel überlegen und danach bewusst ein Zahlentripel nennen, von dem sie glauben, dass es ihrer vorgefertigten Regel nicht gehorcht. Sie müssen also versuchen, ihre Regel zu widerlegen. Antwortet der Versuchsleiter dann dennoch mit „Richtig", können sie sicher sein, dass ihre Regel falsch ist und eine andere Regel konstruieren, die sie auf dem selben Wege testen. Ein mühsames Verfahren, das aber letztendlich als einziges zum Ziel führt.

Laut Wason gingen jedoch ca. 80 % der Probanden nicht so vor, sondern nannten vor allem Zahlentripel, von denen sie glaubten, dass diese ihre Regel bestätigen. Da die obige Regel jedoch relativ einfach ist, erlaubt sie auch viele Aussagen, die im Rahmen von deutlich komplexeren Konstrukten geäußert werden. Daher dauert es bei induktivem Schließen sehr lange, auf den Fehler in der eigenen Regel zu stoßen. Wason argumentiert, dass Personen die fundamentale Tendenz haben, an ihren Annahmen festzuhalten und die Suche nach Informationen zu vermeiden, die ihren Annahmen widersprechen könnten.

Schulisches Prüfen muss daher deduktiv gestaltet werden, das heißt konkret:
- Man nutzt die Doppelverb-Lernziele und definiert daraus die Aspekte, die beobachtet werden sollen.
- Die Fragestellungen werden in einer Form formuliert, die die gewünschten Handlungen anstoßen.
- Die Einzelbeobachtungen dienen dazu, zu überprüfen, ob der zu bewertende Schüler aufgrund der gezeigten Handlungen die erwarteten Fähigkeiten hat – und in welcher Qualitätsstufe.

Allerdings ist auch deduktives Schließen sehr fehleranfällig. Eine erste Fehlerquelle liegt in der Wahl der Einzelbeobachtungen. Diese müssen geeignet sein, die Anfangshypothese sicher zu überprüfen; es müssen also die richtigen Beobachtungen durchgeführt werden. Folgendes Beispiel zeigt einen Fall, in dem die Regel trotz deduktiven Schließens nicht überprüft werden kann:
- Alle Schwarzhaarigen lügen immer. (Regel)
- Sie kennen Jürgen nur per E-Mail. Er behauptet, er sei blond. (Einzelbeobachtung)
- Welche Haarfarbe hat Jürgen?

Die letzte Frage ist mit Hilfe der vorliegenden Beobachtung nicht zu beantworten. Aus der Logik sind verschiedene Konstrukte bekannt, die – meist auf theoretischem Weg – die für individuelle Hypothesenkonstrukte notwendigen Fragestellungen determinieren. Dies führt im Kontext der schulischen Prüfung zu weit. Wichtig ist, dass die Anzahl und Qualität der Beobachtungen eine sichere Überprüfung der Hypothese zulassen muss.

Eine zweite Fehlerquelle des deduktiven Schließens liegt in der menschlichen Tendenz zur Ökonomie. Wir neigen dazu, den leichtesten Weg zu gehen, auch im Denken. Das ist keine schlechte Angewohnheit, sie hat sogar einen unbestreitbaren evolutionären Vorteil, da das Gehirn das energiehungrigste Organ des Menschen ist – in Zeiten knapper Ressourcen ist einfaches Denken also ein echter Überlebensvorteil. Vorschnelle deduktive Schlussfolgerungen führen jedoch oft in die Irre. Ein Beispiel mag dies illustrieren:

In ereignisarmen Zeiten produzieren Medien gerne aus Nebensächlichkeiten Schlagzeilen. So schreibt beispielsweise eine fiktive Boulevardzeitung:

Die Familie ist die Basis eines gesunden Volkes. So steht es jedenfalls in unserer Verfassung. Dass dieser Satz Sinn macht wird aber auch aus folgenden Zahlen deutlich: Verheiratete Männer werden im Schnitt 73 Jahre alt, ledige Männer nur 66 Jahre. Dieser Effekt kann dadurch erklärt werden, dass verheiratete Männer wohl gesünder leben und besser auf sich aufpassen, weil sie ja für Ihre Familie zu sorgen haben. (fiktives Beispiel)

Diese Schlussfolgerung ist schlichtweg falsch, es ist ein rein mathematischer Effekt: Menschen, die im Alter von 0–18 Jahren sterben, sind nur in den seltensten Ausnahmefällen verheiratet. Die verstorbenen Kinder und Jugendlichen werden damit der Gruppe der unverheirateten Menschen zugerechnet und ziehen das arithmetische Mittel dieser Gruppe zu sich heran wie ein Magnet. Rechnet man diesen Effekt heraus, ist der Unterschied nicht mehr bedeutsam.

Hier wurde – aus Bequemlichkeit oder aus anderen Gründen – nicht nach alternativen Beobachtungen gesucht, die eine stabile Prüfung der Hypothese „Verheiratete Männer leben länger als unverheiratete" erlauben.

Die dritte bekannte Fehlerquelle des deduktiven Schließens liegt in der Wahl einer ungeeigneten Hypothese. Das beste Beispiel hierfür ist das bekannte „Ziegenproblem", das hier nur kurz skizziert wird und – bei Interesse – in vielen Publikationen (z. B. von Randow 1999) ausschweifend thematisiert wird:

„Nehmen Sie an, Sie wären in einer Spielshow und hätten die Wahl zwischen drei Toren. Hinter einem der Tore ist ein Auto, hinter den anderen sind Ziegen. Sie wählen ein Tor, sagen wir, Tor Nummer 1, und der Showmaster, der weiß, was hinter den Toren ist, öffnet ein anderes Tor, sagen wir, Nummer 3, hinter dem eine Ziege steht. Er fragt Sie nun: ‚Möchten Sie das Tor Nummer 2?' Ist es von Vorteil, die Wahl des Tores zu ändern?" (Wikipedia 2014).

Die überraschende Lösung lautet: Ja, man sollte wechseln, denn Wechseln verdoppelt die Gewinnchance. Diese Lösung ist zweifelsfrei beweisbar (von Randow, ebd.), dennoch glauben viele Menschen, dass dem nicht so sei, da sich die Situation scheinbar nicht verändert habe. Diese Schlussfolgerung ist falsch, aber gut erklärbar: Aufgrund der Tatsache, dass sich eine Situation entkompliziert, blenden viele Menschen die Vorgeschichte aus, entscheiden sich für die einfache Hypothese – und gelangen so zu falschen Schlussfolgerungen.

Die hier dargestellten Erkenntnisse zu der Problemstellung, auf korrektem Wege Schlussfolgerungen aus Beobachtungen zu ziehen, können für den schulischen Kontext in folgende Regeln überführt werden:

- Jede Beobachtung einer Leistung basiert auf Aufgabenstellungen. Diese können jedoch nicht frei gewählt werden, da solche Einzelbeobachtungen nur induktive Schlüsse zulassen. Die Aufgaben müssen vielmehr so formuliert werden, dass sie es erlauben, eine Hypothese zu prüfen.
- Diese Hypothese wird aus als wahr angenommenen Axiomen abgeleitet und lautet in der Regel: Die Schülerin oder der Schüler hat die Fähigkeit […] wenn sie die Aktion […] zeigt. An einem Beispiel: Eine Schülerin / ein Schüler kann Pflanzen identifizieren, wenn sie mehrere ihr vorgelegte echte Pflanzen korrekt benennt.
- Die Hypothesen sind damit die umformulierten Doppelverb-Lernziele. Dieses Vorgehen ist statthaft, da die Doppelverb-Lernziele – wie beschrieben – aus den vorgegebenen Kompetenzzielen des Lehrplans abgeleitet werden.
- Je nach dem Ergebnis der Einzelbeobachtung(en) kommt man dann zu der stabilen Schlussfolgerung: Die Schülerin oder der Schüler hat die Fähigkeit.

Um jedoch qualitative Abschätzungen über den Stand der jeweiligen Fähigkeit treffen zu können, um also zu erkennen, wie hoch die Fähigkeit ausgeprägt ist, benötigt man ein Instrument, das qualitative Beobachtungen zulässt. Dieses Instrument kann in Form eines Bewertungsrasters generiert werden.

7.3. Bewerten von Leistungsmessungen

Unter der Bewertung versteht man die Einschätzung des Wertes oder der Bedeutung eines Sachverhaltes oder Gegenstandes. Die Bewertung einer Leistungshandlung wird normalerweise anhand eines Bewertungsmaßes oder Bewertungsmaßstabes vorgenommen. In der schulischen Praxis leistet das ein so genanntes Beobachtungsraster (definiert die zu beobachtenden Aspekte) oder ein Bewertungsraster (enthält zusätzlich die Möglichkeit, die einzelnen Aspekte zu bepunkten). Dieses stützt sich dabei eng auf die vorher von jeder Lehrkraft selbstständig erarbeiteten Doppelverb-Lernziele und die Notwendigkeit des deduktiven Schließens. Da Doppelverb-Lernziele von der Lehrperson individuell erstellt werden, kann auch bei Beobachtungsrastern keine allgemeine Gültigkeit postuliert werden, denn das Beobachtungsraster verinnerlicht die Doppelverb-Lernziele und formuliert für jedes dieser Lernziele und für jedes mögliche erbrachte Ausmaß dieser Leistung eine konkrete Aktion. Die schriftliche Fixierung und transparente Vorgehensweise bei der Leistungsbewertung, durch die ein breites Spektrum erbrachter Leistungen sowie deren Bewertung festgehalten wird, dürfte die Unsicherheit auf ein akzeptables Niveau senken.

Ein fertiges Beobachtungsraster enthält somit Fähigkeit, Aktion und mögliche Fehlerquellen, darüber hinaus können optional auch mögliche sehr gute bzw. sehr schlechte Aktionen beschrieben werden, sowie ein dazu passendes handlungsbezogenes Feedback. Für die in Tabelle 5 (S. 00) dargestellte Transferaufgabe zur Berechnung des Flächeninhalts eines Dreieckes könnte ein darauf aufbauendes Beobachtungsraster wie folgt aussehen:

Kog. Leistung	Transfer		
Fähigkeit	Bestimmung des Flächeninhalts eines Dreiecks		
Aktion	Auswahl der geeigneten Formel	Anwendung der Formel	
Fehlerquellen	Die Formel wird nicht genannt oder eine falsche Formel ausgewählt.	Die Werte werden unkorrekt eingesetzt.	Rechenfehler
Gute Aktionen	Die einzelnen Elemente der Formel werden zusätzlich dargestellt/erklärt.	Das Einsetzen der Werte wird begründet dargestellt.	Die einzelnen Schritte der Berechnung werden ausgeführt.
Feedback	Du solltest wissen, welche Formel zu welcher Berechnung gehört.	Versuche in eigenen Worten zu beschreiben, welche Elemente die Formel benötigt. Finde diese Elemente dann in den Angaben zum Dreieck.	Vollziehe die einzelnen Rechenschritte nach (z. B. indem Du sie einer Mitschülerin / einem Mitschüler erklärst).

Tab. 6: Beispiel eines Beobachtungsrasters zur Berechnung des Flächeninhalts eines Dreieckes

Bei der Bewertung geht es – neben der reinen Notengebung – vor allem um die transparente und nachvollziehbare Begründung der Leistungsbewertung und um deren Rückmeldung. Die Herausforderung für die Lehrkraft lautet: Wie formuliere ich meine Rückmeldung so, dass nicht nur die funktionalen Erfordernisse der Leistungsbeurteilung erfüllt werden, sondern auch die pädagogischen Potenziale genutzt werden können?

Das bereits in Kapitel 1 angesprochene Handlungsregulationsmodell von Frese und Zapf (1994) stammt ursprünglich aus der Evaluationsforschung, kann jedoch ohne Probleme für den Bewertungskreislauf der schulischen Leistungsbewertung adaptiert werden. Zusammengefasst ergeben sich aus der Kombination des

Handlungsregulationsmodells mit dem Kontext „Schule“ die folgenden Schlussfolgerungen:

- Die Entscheidung über eine Änderung/Anpassung der Lernhandlungen liegt letztendlich beim Lernenden.
- Die Wahrscheinlichkeit für die Änderung steigt mit der Qualität des Feedbacks.
- Feedback sollte immer so gestaltet werden, dass es der Adressat direkt in Handlungen umsetzen kann.

Je transparenter die Feststellung und Bewertung von Schülerleistungen mittels Note in der Praxis durchgeführt wird, umso nachvollziehbarer ist allen Beteiligten (Schülerinnen und Schülern, Eltern, Lehrerinnen und Lehrern) das Zustandekommen der (endgültigen) Beurteilung. Wenn sich die Maßstäbe der einzelnen Lehrerinnen und Lehrer innerhalb des Fachbereichs oder darüber hinaus sehr unterscheiden, führt dies bei den Beurteilten wie auch bei den Eltern zu einer gewissen Verunsicherung („Warum kriegt meine Tochter bei Frau Meier mit so viel Fehlern eine gute Note und ihre Freundin mit derselben Fehlerzahl bei Frau Müller nur eine ausreichende?“).

Da Schülerinnen und Schüler unabhängig von Lehrperson oder Schulstandort innerhalb eines Schulsystems für ihre Leistungen in vergleichbarer Weise beurteilt werden sollen, ist es erstrebenswert, dass eine größtmögliche Transparenz der Lernziele geschaffen wird. Die Schülerinnen und Schüler müssen von Beginn an wissen, welche Kompetenzen sie erwerben und woran sie diese zeigen sollen. Dies kann durch ein Beobachtungs- oder Bewertungsraster in geeigneter Weise sichergestellt werden. Zudem unterstützt ein geeignetes, selbst erstelltes Raster die Lehrkraft bei der Leistungsbeurteilung – insbesondere in der initialen Berufsphase. Ein Beobachtungsraster nimmt die taxonomisierten Lernziele auf und definiert für jedes Lernziel und jede Leistungsstufe die entsprechende Aktion.

Es sei an dieser Stelle ausdrücklich erwähnt, dass ein Beobachtungsraster keinesfalls den Anspruch auf Vollständigkeit oder didaktische Richtigkeit erhebt. Es eignet sich auch nicht für jedes Anwendungsgebiet, denn je schlechter definiert die Fähigkeiten sind, desto schwieriger ist es, ein Beobachtungsraster zu erstellen.

Bei der Erstellung Ihres Beobachtungsrasters gehen Sie entsprechend den folgenden Punkten vor:

- Strukturieren Sie Ihr Beobachtungsraster nach den Fähigkeiten, die beobachtet werden sollen.
- Erstellen Sie nun zu den jeweiligen Fähigkeiten alle Aktionen, welche die Schülerin bzw. der Schüler zeigen soll, um die Fähigkeit nachzuweisen. Bis hierhin können Sie einfach die Inhalte der erstellten Doppelverb-Lernziele übernehmen.
- Beschreiben Sie nun Fehlerquellen und besonders gelungene Handlungen. Erwarten Sie dabei nicht das Optimum, denn damit steigt die Gefahr von Kontrastfehlern, sondern gehen Sie immer von einer zufriedenstellenden Erwartung aus (vgl. Tab. 6). Dadurch haben Sie die Möglichkeit, besonders guten Leistung mit Beispielen zu belegen und Ihre Schüler/-innen dafür zu loben.
- Empfehlenswert ist es auch, beim Erstellen eines Beobachtungsrasters Ideen für das Feedback an die Schüler/-innen zu sammeln. Das Feedback sollte so gestaltet sein, dass es direkt in Handlungen überführt werden kann. Damit haben Sie bereits einige treffende Vorlagen für die jeweiligen Rückmeldungen (in Form von Randbemerkungen und Zusammenfassung) und verhindern, dass Sie im Eifer des Gefechts wenig (oder zu wenig) handlungsorientierte Rückmeldungen geben (vgl. Tab. 6).

Das folgende Beispiel zeigt die Umsetzung dieser Regeln für die Beobachtung einer Problemerörterung im Fach Deutsch:

Fähigkeit	Aktion	Mögliche Fehlerquellen	Besonders Gelungenes	Feedback
Strukturierung des Aufsatzes	Grobgliederung: Einleitung – Hauptteil – Schluss	• Ein oder mehrere Teile werden vergessen • Falsche Gewichtung der Teile		• Der Schluss ist im Vergleich zum Hauptteil zu kurz. • Du hast die Einleitung vergessen.
Themaerfassung	Formulierung einer dem Thema angemessenen Frage-/Problemstellung	• Einleitung ist zu abstrakt • Erörterungstyp nicht richtig erfasst • Argumente werden schon vorweggenommen • Führt nicht zum Hauptteil hin • Keine Frage-/ Problemstellung • Themaverfehlung • Allgemeinplätze und Floskeln	• Arbeitsweise wird erläutert (Advanced Organizer) • Aktuelle Zahlen, Befunde und Quellen • Besonders treffende Formulierung einer Frage-/Problemstellung • Interesse zum Weiterlesen wird geweckt • Treffend gewähltes Beispiel zur Illustration des Themas	• Prüfe, ob Deine Frage mit einer Erörterung beantwortbar ist • Nutze Beispiele und aktuelle Geschehnisse, um Deine Einführung lebendiger und konkreter zu machen. • Sehr gut finde ich, dass Du Deine Arbeitsweise erläutert hast.
Inhaltsgestaltung	• Einleitung führt zum Thema hin • Passende Argumente • Argumente werden nach Erörterungstyp gegliedert • Reihung der Argumente	• Keine oder ungenaue Überleitungen • Argumente werden wiederholt • keine treffenden Argumente • Begründungen und Beispiele fehlen • unrichtiger Aufbau der Argumente	• viele und gute Beispiele • geschickte Überleitungen • starke und treffende Argumente • gut gewählte Zitate • Quellen und Befunde • Schluss nimmt die Einleitungsfrage auf	• Gute Auswahl der Argumente • Begründe Diene Argumente mit Beispielen/ Quellen/Belegen • Dein Argumentationsgang ist sehr gut strukturiert
	• Begründung der Argumente • Richtiges Verwenden von Zitaten • Geeignete Überleitungen und logische Verknüpfungen • geeignete Beispiele wählen • Zusammenfassung der Argumente und Synthese	• Allgemeinplätze und Floskeln • Das Fazit greift neue Argumente auf • Kein oder unpassendes Fazit • Argumentieren aus eigenen Befindlichkeiten	• Ausblick und weitere Differenzierung des Themas	• Du argumentierst aus eigenen Befindlichkeiten und Überzeugungen

Fähigkeit	Aktion	Mögliche Fehlerquellen	Besonders Gelungenes	Feedback
Sprachliche Gestaltung	• gute Wortwahl, Sprachstil und Grammatik • Text ist leicht und flüssig zu lesen	• Lange und unübersichtliche Sätze • Grammatisch falsch • Kommafehler • Rechtschreibfehler • Umgangs und Szenesprache • Fehler im Ausdruck • Wiederholt gleicher oder falscher Satzbau • Wortwiederholungen • Falsches Verwenden von Fachwörtern / indirekter Rede	• Verwenden von Adjektiven • Verwenden von Fachwörtern • Prägnanz	Fehler der sprachlichen Gestaltung werden im Fließtext angemerkt (Korrekturzeichen)
Äußere Form / äußerer Aufbau	• Absatzgliederung und Strukturierung des Fließtextes • Nummerierung • Schriftbild • Sauberkeit	• Nie eine neue Zeile anfangen • Zu viele Absätze • Unsinnige Absätze • unlesbar • schlechte • Schrift • Tintenkleckse • Sprungmarken und unsaubere Korrekturen	• übersichtliche und sinnzusammenhängende Gliederung • Hierarchische Nummerierung • Sehr schöne Schrift	• Bemühe Dich, Deine u und v unterschiedlich zu gestalten • Benutze ein Lineal zum Durchstreichen • Entwirf Deinen Gedankengang vorher auf einem Notizzettel.

Tab. 7: Beobachtungsraster für eine Problemerörterung im Fach Deutsch

Dieses Beispiel darf jedoch lediglich als Anregung verstanden werden, da Beobachtungsraster, die von außen implementiert werden, meist wenig nützlich sind. Der Grund ist ganz einfach: Jede Lehrkraft konzipiert und gestaltet ihren Unterricht weitgehend autonom, setzt eigene Schwerpunkte und beeinflusst damit auch die passenden Leistungskriterien. Ein von außen übernommenes Beobachtungsraster ist daher nicht hilfreich und kann nur als Impuls dienen. Daher wäre es zu kurz gegriffen, das vorliegende Beobachtungsraster einfach zu übernehmen. Jedes Beobachtungsraster muss adaptiert werden, um es für die selbst gewählten Kriterien nutzbar zu machen.

Becker-Mrotzek und Böttcher (2008) schlagen als geeigneten Basis-Kriterienkatalog für Deutschtests der Sekundarstufe I das Zürcher Textanalyseraster von Nussbaumer und Sieber (1994) vor. Die drei – schon in der antiken Rhetorik geltenden – Grundkriterien Sprache, Inhalt und Aufbau werden zu fünf Dimensionen (Sprachrichtigkeit, Sprachangemessenheit, Inhalt, Aufbau und Schreibprozess) weiterentwickelt,

„die mehr oder weniger für alle Textarten zutreffen. [Der Kriterienkatalog] muss in zweifacher Weise an die jeweils konkreten Schreibaufgaben und Bewertungsziele angepasst werden: Zum einen müssen die spezifischen Merkmale der jeweiligen Textart eingearbeitet werden […]. Und zum anderen müssen die didaktischen Ziele der jeweiligen Unterrichtseinheit berücksichtigt werden. Auf diese Weise kann und soll es zu Verschiebungen in den fünf Dimensionen kommen. So kann es Beispielsweise sinnvoll sein, bei der Beurteilung lyrischer Texte ganz auf die Dimension sprachliche Richtigkeit zu verzichten, während bei argumentativen Texten etwa die Dimensionen Inhalt und Aufbau ausgeweitet werden müssen" (S. 94).

Aus der Angabe des Grades der Einzelleistung (1 = erbracht, 0,5 = teilweise erbracht, 0 = nicht erbracht) lässt sich ein abgestuftes Punkteschema für Klassenarbeiten entwickeln (12 Punkte = Bestleistung).

Dimension	Kriterium	Grad		
		1	0,5	0
	Sprachrichtigkeit			
Orthografie	Entspricht die Orthografie einschließlich Zeichensetzung dem erwarteten Lernstand?			
Grammatikalität	Sind Wortbildung und Satzbau grammatisch korrekt?			
	Sprachangemessenheit			
Wortwahl	Ist der Wortschatz angemessen? Werden Inhaltswörter, Funktionswörter, komplexe Ausdrücke und Fachtermini treffend verwendet?			
Satzbau	Ist der gewählte Satzbau der Aufgabe und dem Leser angemessen?			
	Inhalt			
Gesamtidee	Lässt der Text eine dem Thema angemessene Gesamtidee erkennen?			
Umgang/ Relevanz	Sind Umfang und Inhalt der Aufgabe angemessen?			
	Aufbau			
Textmuster	Wird ein der Aufgabe angemessenes Textmuster verwendet?			
Textaufbau	Ist der Text sinnvoll aufgebaut? Lässt er eine innere/äußere Gliederung erkennen (Abschnitte etc.)?			
Thematische Entfaltung	Wird das Thema in einer der Fragestellungen angemessen entfaltet?			
Leserführung	Wird der Leser aktiv durch den Text geführt? Werden textstrukturierende Mittel verwendet?			
	Prozess			
Planen/ Überarbeiten	Lässt der Text Planungs- und Überarbeitungsspuren erkennen?			
Wagnis/ Kreativität	Lässt der Text ein besonderes sprachliches Wagnis erkennen? Ist er in besonderer Weise kreativ?			

Tab. 8: Kriterienkatalog von Becker-Mrozek und Böttcher (2006, S.95) nach dem Zürcher Textanalyseraster von Nussbaumer und Sieber (1994).

Im Sinne eines respektvollen, lernsentiven Umgangs mit Texten empfehlen Becker-Mrotzek und Böttcher (ebd.) dem Lehrer, eher die Rolle eines Lesers statt eines Beurteilers einzunehmen. Dies schließt eine flexible Anwendung des Kriterienkatalogs ein und meint beispielsweise auch, dieses Bewertungsraster durch die in Tabelle 6 und 7 eingeführten Aspekte *Fehlerquellen, gute Aktionen* und *Feedback* zu ergänzen.

„Maßgeblich für den Bewertungsprozess sind der Schülertext und die Richtung, die er vorgibt, und nicht von

außen an ihn herangetragene Kriterien inhaltlicher oder sprachlicher Art. Dies erscheint zunächst widersprüchlich zu der von uns vorgestellten Arbeit mit Kriterienkatalogen. Tatsächlich ist aber die Forderung und Notwendigkeit, Schülertexte als Texte ernst zu nehmen, die Basis für alle […] Korrekturkonzepte" (S. 96).

Das Erstellen eines Beobachtungs- oder Bewertungsrasters in Kombination mit der Beurteilung nach dem BIB-Modell hilft nicht nur, Sicherheit in das eigene Handeln zu bringen, sondern auch Transparenz in dem von Schüler/-innen und Eltern so oft als „undurchsichtige Notengebung" empfundenen Verfahren herzustellen. Man erlangt dank der Fundierung durch die Ziele des Lehrplans und dank des selbst erstellten Rasters Routine und Gradlinigkeit in der Aufsatzkorrektur. Da dies ein langjähriger Prozess ist, stellen sich die Früchte dieser Arbeit allerdings nicht unmittelbar nach Implementierung dieser Vorgehensweise ein, sondern benötigen Zeit und reflexive Erfahrung.

7.4. Bilanzierung des Gelernten

In diesem Kapitel werden folgende Lernimpulse gesetzt:

- Lehrkräfte sind meist gut darin, Unterschiede innerhalb einer Klasse richtig darzustellen. Sie haben allerdings Probleme, die Leistung klassenübergreifend einzuordnen. Daher ist es sinnvoll, verschiedene Elemente der Leistungsdiagnostik in ihr Lehrerhandeln einzubinden.
- Die erwarteten Leistungen müssen für die Lernenden transparent sein. Dies gelingt durch die Formulierung klarer Lernziele, z. B. durch Verwendung von Wissenstaxonomien, sowie durch die Operationalisierung erwarteter Handlungen, z. B. durch Doppelverb-Lernziele.
- Die Beobachtungen müssen möglichst neutral interpretiert werden. Dazu hilft ein deduktives Vorgehen, bei dem Schülerhandeln aus Doppellernzielen abgeleitet wird, mehr, als ein induktives Vorgehen, bei dem die Lehrkraft einfach fragt, um daraus auf das eventuelle Wissen der Schüler/-innen zu schließen. Allerdings muss auch ein deduktives Schließen gewissen Qualitätsstandards entsprechen.
- Auch die Bewertung der Leistungen muss transparent sein. Das gelingt am besten mit Kriterienkatalogen, die auf den Doppellernzielen aufbauen, häufige Fehlerquellen aufzeigen und diese in Handlungsanweisungen umsetzen.

Zuletzt folgt wieder ein Blick auf die Episoden, mit denen Sie am Anfang des Kapitels konfrontiert wurden. Dem Referendar aus Episode 1 sollten Sie mit Hilfe der in diesem Kapitel aufgezeigten Elemente jetzt problemlos Tipps geben können, wie man einen Deutschaufsatz objektiv und nachvollziehbar korrigiert.

Bei der Schülerin aus Episode 2 haben Sie als Referendar sicherlich mit gutem Willen gehandelt. Sie haben Ihre Korrektur nicht nur nachvollziehbar und transparent gemacht, sondern haben sie auch noch nach verschiedenen Kriterien aufgespaltet. Allerdings zeigt die Reaktion der Schülerin, dass dieser Versuch von der Adressatin nicht richtig verstanden wurde. Von daher wäre der Nachvollzug der Korrektur besser mit Hilfe der hier vorgestellten Maßnahmen umzusetzen.

Was den Kollegen aus Episode 3 und seine Abfragetipps angeht, so sollten Sie eigentlich schon aus dem letzten Kapitel wissen, dass man es sich nicht ganz so einfach machen sollte. Eine solche Abfrage ist weder objektiv noch reliabel oder valide.

Das letzte Kapitel blickt nun auf Situationen, in denen diese Form der Leistungsbeurteilung zu kurz greift: Offene Unterrichtsszenarien oder Projekte. Hier müssen weiterführende Bewertungsaspekte gefunden werden, um auch in diesen Situationen die Kompetenz der Lernenden einzuschätzen.

8. Jenseits der Standardisierung – ein Ausblick

Setzten die letzten Kapitel den Schwerpunkt eher auf den vom Schulrecht definierten Normalvorgang der Bewertung, wird an dieser Stelle ein Ausblick auf jene Anforderungen der schulischen Leistungsbewertung gegeben, die nicht so weit definier- und standardisierbar sind. Neben der im Kapitel 3 angesprochenen Förderung geht es hier vor allem um die Bewertung von Handlungen im Unterricht, die eine übliche Form der Urteilsbildung nicht zulassen: Offene Umgebungen, Gruppenarbeit, Projekte oder die Beurteilung von Lernprozessen (Winter 2004).

All diesen Situationen ist gemein, dass es bei ihnen mit deutlich mehr Anstrengungen verbunden ist, die im letzten Kapitel beschriebenen Vorgänge umzusetzen. Grundsätzlich gilt: Je schwerer es ist, Leistungsanforderungen konkret zu beschreiben, desto komplexer ist auch der damit verbundene Bewertungsvorgang. Gerade das Beispiel der offenen Unterrichtsszenarien zeigt deutlich, dass es hier eine Menge mehr Aspekte im Bewertungsprozess zu beachten gibt (und geben muss) als bei der Korrektur schriftlicher Arbeiten oder bei der mündlichen Abfrage. Der notwendige Aufwand potenziert sich dadurch.

Hinzu kommt die Einschränkung, dass es einem einzelnen Bewerter nur sehr schwer möglich ist, mehrere Aspekte gleichzeitig im Auge zu behalten. Hier gilt sinngemäß das Gesetz von George A. Miller (1956), wonach Menschen nicht in der Lage sind, eine Variable in mehr als sieben Qualitätsabstufungen zu beschreiben und mit diesen Beschreibungen zu operieren (auch das könnte implizit ein Grund für Notensysteme sein, die mit weniger Stufen auskommen). Gleiches gilt auch für mehrdimensionale Aspekte. Miller beschreibt hierzu verschiedene Experimente, die sehr gut zeigen, dass ab einer gewissen Multidimensionalität einzelner Aspekte keine eindeutigen Differenzierungen in der Beurteilung der verschiedenen Dimensionen mehr getroffen werden können. Übertragen auf die Bewertung von Gruppenarbeiten hieße das beispielsweise: Wenn eine einzelne Lehrkraft den Arbeitsprozess einer Gruppe mit vier Mitgliedern in den Dimensionen *Koordination, Wissenskommunikation, Qualität der Inhalte* und *Argumentationsgang* in nur vier Abstufungen bewerten soll, wird es einer unerfahrenen Lehrperson nicht mehr möglich sein, das problemlos umzusetzen.

Diese Herausforderung kann durch zwei Maßnahmen gemeistert werden:

- Reduzierung der Dimensionen oder der Abstufungen, was die Qualität der Bewertung so weit schmälert, dass ein qualitatives Feedback nicht mehr gegeben werden kann oder
- der Erwerb von Expertise durch die reflexive, theoriegeleitete Auseinandersetzung mit Erfahrungen, die zu professionellen Kompetenzen einer Lehrperson führen. Denn, so schreibt Miller, „Now we are beginning to get up into the range that ordinary experience would lead us to expect“ (S. 90).

Erkenntnisse der Expertiseforschung konnten den Nachweis erbringen, dass es Experten deutlich besser als Novizen möglich ist, Muster von mehrdimensionalen Variablen zu erkennen und zu nutzen. Bekannt ist hier beispielsweise das Experiment von Charness (1981), der Schachexperten und -novizen verschiedene Stellungen jeweils nur wenige Sekunden lang präsentierte. Im Ergebnis konnten Experten deutlich mehr Schachstellungen fehlerfrei reproduzieren, wohingegen sich ein Schachnovize nur an wenige Figuren erinnern konnte. Experten, so erklären es Gruber und Ziegler (1997), haben die Fähigkeit, Wissen prinzipienbasiert abzuspeichern und können sich damit vielfältiger elaborierter, fallbasierter Wissensschemata (sogenannter Chunks) zur funktionellen Einordnung neuer Information bedienen.

Auf die Bewertung von offenen Unterrichtsszenarien übertragen führt das zu der Empfehlung, sich in einem längerfristigen Prozess mit den Aspekten der Beurteilung offener und damit komplexer und multiperspektivischer Szenarien reflexiv auseinanderzusetzen. Erfahrungen sind dabei als Grundlage kompetenten Handelns wichtig, genügen aber nicht, wenn man daraus keine Handlungen ableitet.

Hilfreich sind dazu die Arbeiten von Thorsten Bohl (2006), der den Versuch unternimmt, Bewertungskriterien für problemzentrierte Unterrichtsszenarien zu erstellen, um die von ihm beschriebene Kluft zwischen offenen Unterrichtsszenarien und Leistungsbeurteilung zu überbrücken. Dabei setzt er seinen Fokus nicht so sehr auf festgelegte Kriterien, die lediglich die Bewertungsdimensionen in für den Laien nicht mehr fassbare Größenordnungen transportieren würden, sondern gibt gelungene Anregungen für eine Verknüpfung der auf den ersten Blick nicht zusammenpassenden Aspekte *Multidimensionalität* und *Exaktheit der Beurteilung* und erschafft dadurch Bewertungsschemata, die einen einfacheren Zugang in den Erwerb von Beurteilungskompetenz erlauben.

Bei einer auf diese Weise durchgeführten Leistungsbewertung stehen weniger die Selektion (und damit die extrinsische Motivation) als vielmehr Hilfestellungen für und die Förderung der Lernenden an erster Stelle. Die von ihm erstellten Ansätze sind in ihrer Genauigkeit und Durchführbarkeit nicht mit den im letzten Kapitel vorgestellten Bewertungsrastern zu vergleichen und erfordern von der Lehrkraft ein sehr hohes Maß an diagnostischer Kompetenz. Dennoch geben sie Anregungen für die Gestaltung von Bewertungskriterien in nicht-strukturierbaren Aufgabenfeldern und sind daher als gute Empfehlung für den erweiterten Kompetenzerwerb in diesem Bereich zu sehen. Dies gilt umso mehr, da es, im Gegensatz zur klassischen Beurteilung, noch keine hohen Erfahrungswerte für offene Unterrichtsszenarien gibt:

„Während Freiarbeit, Arbeit nach Wochenplan, projektorientiertes Arbeiten u. a. en vogue sind, hinken die bildungspolitisch und systemisch ermöglichten Bewertungsverfahren hinterher […. So kommt es], dass die Rezeption der Bewertungspraxis der Reformpädagogik bisher weitgehend vernachlässigt wurde. Dies hatte zur Folge, dass in der für die Lehreraus- und -fortbildung maßgeblichen und anerkannten Literatur reformpädagogische Unterrichtsmodelle zwar vermittelt wurden, allerdings verkürzt um die Bewertungsverfahren […] Die Vernachlässigung von Bewertungsverfahren ist daher zwar verständlich, gleichwohl schulpraktisch ungelöst: Die Unterrichtsmodelle sollten angewandt werden, die Bewertungsmodelle (sofern sie vorhanden und bekannt waren) konnten nicht angewandt werden, allenfalls in Bruchstücken" (Bohl 2006, S. 56 f.).

Das legt die Konsequenz nahe, dass aufgrund fehlender Bewertungsrichtlinien dem einzelnen Lehrenden ein hohes Maß an Eigenaktivität zukommt, will auch er den offenen Unterricht als Grundlage für die individuelle Förderung nehmen – was in diesem Zusammenhang selbstverständlich erscheint. Bohl (ebd.) liefert eine Reihe von Vorschlägen, um die eigene Bewertungspraxis zu optimieren, wobei er gleichzeitig darauf hinweist, dass die Vielzahl der unterschiedlichen didaktischen Möglichkeiten auch eine Vielfalt der Beurteilungskonzeptionen verlangt. Bei der Umsetzung schlägt er einen ähnlichen Weg vor, wie er im letzten Kapitel angedacht wird:

- Aufbau eines Kriterienkatalogs,
- Festlegen des Beobachtungsfokusses durch Beispiele,
- Einordnen der Handlungsqualitäten in ein Bewertungsraster.

Der erste Schritt ist damit die Sammlung von Bewertungskriterien. Hier ein Beispiel für Projektarbeiten:

Lernbereiche	Beispiele für konkrete Bewertungskriterien „Der Schüler / die Schülerin ist in der Lage …
Informations-beschaffung	… im Internet gezielt zu recherchieren." … in Bibliotheken gezielt zu recherchieren." … Quellenangaben korrekt zu benennen." …
Arbeitstechniken	… mehrseitige Texte zusammenfassen." … mit Lexika und Wörterbüchern effektiv zu arbeiten." … einen (Heft-)Aufschrieb sauber und strukturiert zu erstellen." … ein aussagekräftiges Verlaufsprotokoll zu erstellen." …
Selbstständiges Arbeiten	… einen Zeit- und Arbeitsplan für ein Projekt zu erstellen." … den Zeit- und Arbeitsplan für ein Projekt selbstständig zu kontrollieren." … den Arbeitsprozess rückblickend zu analysieren." … Aufgaben in freien Arbeitsphasen selbstständig zu bearbeiten." … Lösungen selbstständig zu kontrollieren." …

Visualisierungs-techniken	… Mindmapping zu erstellen und gezielt einzusetzen." … Strukturlegepläne zu erstellen und gezielt einzusetzen." … Metapläne zu erstellen und gezielt einzusetzen." … Grafiken zu erstellen und gezielt einzusetzen." …
Präsentations-techniken	… eine Präsentation zu strukturieren." den Beginn einer Präsentation adressatenspezifisch zu gestalten." … das Ende der Präsentation adressatenspezifisch zu gestalten." … Medien (Overhead, Powerpoint) gezielt einzusetzen." …
Kooperations-fähigkeiten	… Beratung aktiv und gezielt einzufordern." … selbst Hilfe zu geben." verschiedene Gruppenfunktionen (Zeitwächter, Fahrplanwächter) einzunehmen." …
Kommunikative Aspekte	… Argumente überzeugend vorzutragen." … einen freien Vortrag von fünf Minuten zu halten." … Gesprächsverhalten zu analysieren." … zuzuhören und andere Ansichten aufzugreifen." …

Tabelle 9: Sammlung von Bewertungskriterien für eine Projektarbeit. Quelle: Bohl 2006, S. 97.

Bohl erkennt in diesem Zusammenhang die notwendige Vielschichtigkeit möglicher Bewertungskriterien an und ermuntert dazu, die notwendigen beobachtbaren Handlungen unter Bezugnahme auf die Klassenzusammensetzung und vor allem auf die eigene Berufsexpertise abzuleiten. Zwar operiert Bohl nicht mit Doppelverb-Lernzielen, schlägt aber vor, den Beobachtungsfokus durch vielfältige Handlungsbeispiele zu festzulegen, was im Ergebnis auf das Gleiche hinausläuft:

„Das Bewertungskriterium „Der Schüler / die Schülerin ist in der Lage, Medien (z. B. Overhead) gezielt einzusetzen". Hier wären drei Aspekte zu unterscheiden: zum einen die Gestaltung der eingesetzten Folien (mögliche Indikatoren: Lesbarkeit, Übersicht, Aussagekraft, Einsatz von Symbolen, Farben, Strukturierungshilfen, kreative Elemente), zum Zweiten der Einsatz der Folien (mögliche Indikatoren: sinnvolles Abdecken, Lesbarkeit und Schärfe überprüfen, Einsatz im Präsentationsverlauf). Jede Folie sollte zudem einen Bezug zum Thema haben. Im Unterricht könnten anhand qualitativ unterschiedlicher Folien wesentliche Merkmale und Anwendungsfehler herausgearbeitet und schriftlich fixiert werden" (Bohl ebd., S. 96).

Solche Tabellen im Schulalltag anzuwenden, Handlungsmuster zu erkennen und Lerngruppen zu bewerten bedarf längerer wohldurchdachter Übung, um die von Miller (1956) beschriebene kognitive Eingeschränktheit durch Expertise zu überwinden. Ebenso kann man aus Ökonomiegründen nicht jede einzelne Abfrage auf diesem Niveau ausarbeiten, aber es erscheint doch sinnvoll, sich für Situationen, die im Beurteilungsalltag immer wiederkehren, ein Repertoire an derartigen Beurteilungsinstrumenten zuzulegen. Man wird natürlich nicht von Anfang an für jede Beurteilungssituation derartig ausgefeilte Instrumente zur Hand haben. Die stetige Erweiterung des Beurteilungsrepertoirs, die Reflektionen der eigenen Bewertungserfahrungen und die dadurch angestoßenen Verbesserungen in Ihrer Art, Schülerinnen und Schüler zu bewerten, werden Ihnen mit der Zeit helfen, professionell, transparent, gerecht, kurz gesagt gut in Ihren Beurteilungen zu werden.

So endet dieses Buch, wie es begann: Mit einem Appell, den Lehrberuf als Profession zu sehen, die nur durch reichhaltige, elaborierte Erfahrungen im Handlungskontext erworben werden kann. Die notwendigen theoretischen Grundlagen der Leistungsbewertung sind vorhanden, bilden aber in erster Linie die Reflexionsebene für die eigene Handlungspräzisierung. Dennoch lohnt es sich, diesen Weg zu wählen – zum Nutzen aller Beteiligten. Denn nur so

verlieren die in den einzelnen Kapiteln geschilderten Episoden und noch viele weitere mögliche und unmögliche Ereignisse aus der Schulwirklichkeit für Sie den Status der Bedrohung und werden zu Herausforderungen für Ihr professionelles und kompetentes Lehrerhandeln. Viel Erfolg!

Literatur

Arnold, K.-H. (2002): Qualitätskriterien für die standardisierte Messung von Schulleistungen. In: F. E. Weinert (Hrsg.), Leistungsmessungen in Schulen. Weinheim: Beltz, 117–130.

Bauer, K.-O. / Kopka, A. / Brindt, S. (1996): Pädagogische Professionalität und Lehrerarbeit. Eine qualitativ empirische Studie über professionelles Handeln und Bewußtsein. Weinheim/München: Juventa.

Bayerische Landesstelle für den Schulsport (2014): Empfehlungen zur Leistungsbewertung im Fach Sport für die weiterführenden Schulen in Bayern. München: Bayerische Landesstelle für den Schulsport.

Bayerisches Staatsministerium für Bildung und Kultus, Wissenschaft und Kunst (StMUK) (2012): Schule in Bayern. Die rechtlichen Grundlagen. Zugriff am 16. 11. 2012 unter http://www.km.bayern.de/eltern/was-tun-bei/rechte-und-pflichten.html

Becker-Mrotzek, M. / Böttcher, I. (2006): Schreibkompetenz entwickeln und beurteilen, Praxishandbuch für die Sekundarstufe I und II. Berlin: Cornelsen Scriptor.

Berliner, D. C. (2001): Learning about and learning from expert teachers. In: International journal of educational research, Jg. 35, Heft 5, 463–482.

Bildungsdirektion Kanton Zürich (2013): Beurteilung und Schullaufbahnentscheide. Über das Fördern, Notengeben und Zuteilen. Zürich: Lehrmittelverlag des Kantons Zürich.

Bloom, B. S. (1972): Taxonomie von Lernzielen im kognitiven Bereich. Weinheim: Beltz.

Bohl, T. (2006): Prüfen und Bewerten im Offenen Unterricht. Weinheim: Beltz.

Bortz, J. (1995): Lehrbuch der empirischen Forschung. Berlin: Springer.

Bortz, J. / Döring, N. (1995): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. Berlin: Springer.

Bromme, R. (1992): Der Lehrer als Experte. Zur Psychologie des professionellen Wissens. Bern: Huber.

Bügelmann, H. / Backhaus, A. / Brinkmann, E. / Coelen, H. / Franzkowiak, T. / Knorre, S. / Müller-Naendrup, B. / Oser, E. / Roth, S. (2006): Sind Noten nützlich – und nötig? Ziffernzensuren und ihre Alternativen im empirischen Vergleich. Siegen: Universität Siegen, Arbeitsgruppe Primarstufe.

Bundesvereinigung der Deutschen Arbeitgeberverbände (1998): Schule in der modernen Leistungsgesellschaft. Positionspapier des BDA.

Campbell, D. T. (1976/2011): Assessing the Impact of Planned Social Change. Reprint von 1976. In: Journal of MultiDisciplinary Evaluation, Jg. 7, Heft 15, 3–43.

Chaikin, A. / Sigler, E. / Derlega, V. (1974): Nonverbal Mediators of Teacher Expectancy Effects. In: Journal of Personality and Social Psychology, Jg. 30, Heft 1, 144–149.

Charness, N. (1981): Aging and skilled problem solving. In: Journal of Experimental Psychology: General, Heft 110 (1), 21–38.

Cohen, E. G. (1994): Designing groupwork. Strategies for the Heterogeneous Classroom. New York: Teachers Colleg Press.

Collins, A. / Brown, J. S. / Newman, S. E. (1989): Cognitive apprenticeship: Teaching the craft of reading, writing and matematics. In: L. B. Resnick (Ed.), Knowing, learning and instruction. Essays in honor of Robert Glaser. Hillsdale, NJ: Erlbaum, 453–494.

Cornelsen-Akademie (2011): Täuschungsversuch, Abschreiben lassen. In: Best of Schulrecht. Schulrechtsfälle zum Thema Leistungsbewertung. Berlin: Cornelsen-Akademie.

Deci, E. L. / Ryan, R. M. (1993): Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. In: Zeitschrift für Pädagogik, Jg. 39, Heft 2, 223–238.

Deci, E. L. / Ryan, R. M. (2000): The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. In: Psychological Inquiry 2000, Jg. 11, Heft 4, 227–268.

Donoghue, F. (2010): Wo viel Licht ist, ist auch viel Schatten. Hochschulen in den USA zwischen Elite und Titelmühle. In: Forschung & Lehre, Jg. 17, Heft 9/10, 632–635.

Dreyfus, H. L. / Dreyfus, S. E. (1986): Mind over machine: The power of human intuition and expertise in the era of the computer. New York: The Free Press.

Endruweit, G. / Trommsdorf, G. (1989). Wörterbuch der Soziologie. Stuttgart: Enke. Zugriff am 14. 11. 2012 unter https://www.hf.uni-koeln.de/data/ eso/File/seibel/LG70.pdf

Ericsson, K. A. / Krampe, R. T. / Tesch-Römer, C. (1993): The Role of Deliberate Practice in the Acquisition

of Expert Performance. In: Psychological Review, Jg. 100, Heft 3, 363–406.
Fend, H. (1980): Theorie der Schule. München: Urban & Schwarzenberg.
Festinger, L. (1957): A Theory of Cognitive Dissonance. Stanford, CA: Stanford University Press.
Frese, M. / Zapf, D. (1994): Action as the core of work psychology: A German approach. In: H. C.Triandis / M. D. Dunnette / L. M. Hough (Eds.), Handbook of industrial and organizational psychology (Vol. 4). Palo Alto, CA: Consulting Psychologists Press, 271–340.
Gagné, R. / Briggs, L. / Wager, W. (1992): Principles of Instructional Design. Fort Worth, TX: HBJ College Publishers.
Galton, F. (1899): Natural Inheritance. London: MacMillan & Co. Faksimile unter http://galton.org/books/natural-inheritance/pdf/galton-nat-inh- 1up-clean.pdf
Giesecke, H. (2007): Pädagogik als Beruf. Grundformen pädagogischen Handelns. Weinheim: Juventa.
Glasersfeld, E. v. (1992): Konstruktion der Wirklichkeit und der Begriff der Objektivität. In: H. Gumin / H. Meier (Hrsg.), Einführung in den Konstruktivismus. München: Piper, 9–39.
Glasersfeld, E. v. (1996): Der Radikale Konstruktivismus, Frankfurt a. M.: Suhrkamp.
Glasersfeld, E. v. (2006): Einführung in den radikalen Konstruktivismus. In: P. Watzlawick (Hrsg.), Die erfundene Wirklichkeit. München: Piper, 16–38.
Glöckel, H. (1996): Vom Unterricht. Bad Heilbrunn: Klinkhardt.
Gruber H. (1994): Expertise. Modelle und empirische Untersuchungen. Westdeutscher Verlag: Opladen.
Gruber, H. / Stöger, H. (2011): Experten-Novizen-Paradigma. In E. Kiel / K. Zierer (Hrsg.), Basiswissen Unterrichtsgestaltung. Band 2: Unterrichtsgestaltung als Gegenstand der Wissenschaft. Baltmannsweiler: Schneider Verlag Hohengehren, 247–264.
Gruber, H. / Ziegler, A. (1997): Deliberate practice among chess players. In: Sportonomics, Heft 3, 55–61.
Hacker, W. (2006): Psychische Regulation „geistiger" Tätigkeiten: Denkhandeln? In: P. Sachse / W. G. Weber (Hrsg.), Zur Psychologie der Tätigkeit. Bern: Huber, 13–28.
Hartig, J. / Klieme, E. (2007): Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Bonn, Berlin: Bundesministerium für Bildung und Forschung.
Hattie, J. (2009): Visible Learning. A Synthesis of over 800 Meta-Analyses Relating to Achievement. New York, NY: Routledge.
Hattie, J. (2013): Lernen sichtbar machen. Überarbeitete deutschsprachige Ausgabe von Visible Learning. Übersetzt und überarbeitet von Wolfgang Beywl und Klaus Zierer. Baltmannsweiler: Schneider.
Hattie, J. (2014): Lernen sichtbar machen für Lehrpersonen. Baltmannsweiler: Schneider.
Heller, K. A. / Hany, E. A. (2002): Standardisierte Schulleistungsmessungen. In: F. E. Weinert (Hrsg.), Leistungsmessungen in Schulen. Weinheim: Beltz, 87–102.
Helmke, A. (2004): Unterrichtsqualität erfassen, bewerten, verbessern. Seelze: Kallmeyersche Verlagsbuchhandlung.
Helsper, W. (1996): Antinomien des Lehrerhandelns in modernisierten pädagogischen Kulturen. Paradoxe Verwendungsweisen von Autonomie und Selbstverantwortlichkeit. In: A. Combe / W. Helsper (Hrsg.), Pädagogische Professionalität. Untersuchungen zum Typus pädagogischen Handelns. Frankfurt a. M.: Suhrkamp, S. 521–569.
Herbart, J. F. (1986): Systematische Pädagogik. Eingeleitet, ausgewählt und interpretiert von D. Benner. Stuttgart: Klett-Cotta.
Huberman, M. (1980): Das Selbstkonzept. Eine Untersuchung über die Wirkung von Noten, Ranglisten und Preisen auf Kinder der Genfer Primarschule. Genève: FAPSE.
Ingenkamp, K. (1971): Die Fragwürdigkeit der Zensurengebung. Weinheim: Beltz.
Ingenkamp, K. / Lissmann, U. (2008): Lehrbuch der Pädagogischen Diagnostik. Weinheim: Beltz.
Jachmann, M. (2003): Noten oder Berichte? Die schulische Beurteilungspraxis aus der Sicht von Schülern, Lehrern und Eltern. Opladen: Leske + Budrich.
Jacob, B. A. / Levitt, S. D. (2004): To Catch a Cheat. Education Next. Zugriff am 02. 11. 2010 unter http://pricetheory.uchicago.edu/levitt/Papers/ JackobLevittToCatchACheat2004.pdf
Jakobs, B. (2010): Gütekriterien von Noten. Zugriff am 28. 09. 2010 unter http://www.phil.uni-sb.de/~jakobs/paedpsych/noten/guetekriterien_ von_noten.html
Jerusalem, M. / Mittag, W. (1999): Selbstwirksamkeit, Bezugsnormorientierung, Leistung und Wohlbefinden

in der Schule. In: M. Jerusalem / R. Pekrun (Hrsg.), Emotion, Motivation und Leistung. Göttingen: Hogrefe, 223–245.

Johnson-Laird, P. N. / Byrne, R. M. J. (1991): Deduction. Hove, UK, Hillsdale, NJ: Lawrence Erlbaum Associates.

Jürgens, E. (2005): Leistung und Beurteilung in der Schule. Sankt Augustin: Academia.

Keller-Schneider, M. (2010): Entwicklungsaufgaben im Berufseinstieg von Lehrpersonen. Beanspruchung durch berufliche Herausforderungen im Zusammenhang mit Kontext- und Persönlichkeitsmerkmalen. Münster: Waxmann.

Kiel, E. (2001): Grundstrukturen wissenschaftlicher Diskurstätigkeit. Beschreiben, Interpretieren, Bewerten, Erklären, Begründen, Beweisen, Rechtfertigen, Bestreiten. In: T. Hug (Hrsg.), Wie kommt Wissenschaft zu Wissen, Band 1: Einführung in das wissenschaftliche Arbeiten. Baltmannsweiler: Schneider Verlag Hohengehren, 56–68.

Kiel, E. (2007): Epistemologie pädagogischen Handelns. In: G. Reinmann / J. Kahlert (Hrsg.), Der Nutzen wird vertagt … Bildungswissenschaften im Spannungsfeld zwischen wissenschaftlicher Profilbildung und Nutzenorientierung. Lengerich: Pabst, 46–63.

Knauf, M. (2005): Deduktion, logisches Denken. In: J. Funke (Hrsg.), Enzyklopädie der Psychologie. Band C/II/8: Denken und Problemlösen. Göttingen: Hogrefe, 167–264.

Körber, S. (2001): Leistungsbeurteilung. München: Grin.

Kornadt, H.-J. (1978): Abschlussbericht über die wissenschaftliche Begleituntersuchung zum Schulversuch Oberstufe Saar 1970–1977. Saarbrücken: Universität des Saarlandes.

Krampen, G. (1987): Effekte von Lehrerkommentaren zu Noten bei Schülern. In: R. Olechowski / E. Persy (Hrsg.), Fördernde Leistungsbeurteilung. Ein Symposion. Wien: Jugend und Volk, 207–227.

Krathwol , D. R. / Bloom, B. S. / Masia, B. B. (1975): Taxonomie von Lernzielen im affektiven Bereich. Weinheim, Belz.

Kretschmer, H. / Stary, J. (2007): Beobachten in der Schule. In: H. Kretschmer / J. Stary (Hrsg), Schulpraktikum. Eine Orientierungshilfe zum Lernen und Lehren. Berlin: Cornelsen Verlag Scriptor, 19–40.

Lerche, T. (2009): Lernen muss man immer noch selbst! In: U. Dittle / J. Krameritsch / N. Nistor / C. Schwarz / A. Thillosen (Hrsg.), E-Learning: Eine Zwischenbilanz. Kritischer Rückblick als Basis eines Aufbruchs. Münster: Waxmann, 165–178.

Lerche, T. (2012): Übung. In: E. Kiel (Hrsg.), Unterricht sehen, analysieren, gestalten (2. Auflage). Bad Heilbrunn: Klinkhardt, 143–169.

Levitin, D. J. (2006): This Is Your Brain On Music. The Science of a Human Obsession. New York: Dutton/Penguin.

Lind, G. (2008): Amerika als Vorbild? Erwünschte und unerwünschte Folgen aus Evaluationen. In: T. Bohl/H. Kiper (Hrsg.), Lernen aus Evaluationsergebnissen – Verbesserungen planen und implementieren. Bad Heilbrunn: Klinkhardt, 78–97.

Mager, R. F. (1994): Lernziele und Unterricht (Aus dem Amerikanischen übertragen von Helga Thomas). Weinheim: Beltz.

Merrill, D. (2002): First Principles of Instruction. In: Educational Technology, Research and Development, Jg. 50, Heft 3, 43–59.

Miller, G. A. (1956): The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. In: The Psychological Review, Heft 63, 81–97.

Moosburger, H. / Kelava, A. (2007): Testtheorie und Fragebogenkonstruktion. Berlin: Springer.

Nerdinger, F. / Blickle, G. / Schaper, N. (2008): Lehrbuch Arbeits- und Organisationspsychologie. Heidelberg, Berlin, New York: Springer.

Neuweg, H. G. (2008): Qualitätsentwicklung und -sicherung in der schulischen Leistungsbeurteilung – Maßnahmenbereiche und Entwicklungsziele. In U. Stadler-Altmann / J. Schindele / A. Schraut (Hrsg.), Neue Lernkultur – neue Leistungskultur. Bad Heilbrunn: Klinkhardt, 304–322.

Nichols, S. L. / Berliner, D. C. (2005): The Inevitable Corruption of Indicators and Educators Through High-Stakes Testing. San Antonio, TX: NEPC Legacy Publication.

Nichols, S. L. / Glass, G.V. / Berliner, D. C. (2006): High-Stakes Testing and Student Achievement: Does

Accountability Pressure Increase Student Learning? In: education policy analysis archives, Jg. 14, Heft 1, Zugriff am 02. 11. 2010 unter http://epaa.asu.edu/epaa/v14n1/

Nussbaumer, M. / Sieber, P. (1994): Texte analysieren mit dem Zürcher Textanalyseraster. In: P. Sieber (Hrsg.), Sprachfähigkeiten – Besser als ihr Ruf und nötiger denn je! Aarau: Verlag Sauerländer, 141–186.

Oelkers, J. (2002): Leistungen und Noten: Probleme der Schülerbeurteilung. Vortrag anlässlich der Fortbildungstagung des Gymnasiums Hofwil im coop-Zentrum Muttenz am 11.Februar 2002.

Oelkers, J. (2007): Auswirkungen bildungspolitischer Veränderungen auf die Arbeit der Gymnasien und Konsequenzen für die Leitung. Vortrag auf der Tagung „Schulmanagement in Gymnasien – Bildungsprozesse gestalten" am 16. November 2007 in der Europäischen Akademie Schleswig-Holstein, Sankelmark.

Piaget, J. (1972): Die Entwicklung des Erkennens (Bd. 1. Das mathematische Denken). Stuttgart: Klett-Cotta.

Piaget, J. (1981): Einführung in die genetische Erkenntnistheorie (2. Auflage): Frankfurt a. M.: Suhrkamp.

Popper, K. (1935/2005): Logik der Forschung. 11. Auflage. Tübingen: Mohr/ Siebeck.

Prenzel, M. (1997): Sechs Möglichkeiten, Lernende zu demotivieren. In: H. Gruber / A. Renkl (Hrsg.), Wege zum Können. Determinanten des Kompetenzerwerbs. Bern: Huber, 32–44.

Randow, G. v. (1999): Das Ziegenproblem. Denken in Wahrscheinlichkeiten. Stuttgart: rororo.

Rheinberg, F. (1980): Leistungsbewertung und Lernmotivation. Göttingen: Hogrefe.

Rheinberg, F. (2002): Bezugsnormen und Schulische Leistungsbeurteilung. In: F. E. Weinert (Hrsg.), Leistungsmessungen in Schulen. Weinheim: Beltz, 59 – 71.

Rosemann, B. (1975): Prognosemodell für die Schullaufbahnberatung. Ein methodologischer Beitrag zur Bildungsberatung. In: K. Heller (Hrsg.), Handbuch der Bildungsberatung, Band 2. Stuttgart: Klett, 429–447.

Rosemann, B. / Bielski, S. (2001): Pädagogische Psychologie. Weinheim: Beltz.

Rosenthal, R. / Jacobson, L. (1966): Teachers' expectancies: Determinants of pupils' IQ gains. In: Psychological Reports, Jg. 19, 115–118.

Schlömerkemper, J. (2002): Leistungsmessung und Professionalisierung des Lehrberufs. In: F. E. Weinert (Hrsg.), Leistungsmessungen in Schulen. Weinheim: Beltz, 311–321.

Schrader, F. W. / Helmke, A. (2001): Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Hrsg.), Leistungsmessungen in Schulen. Weinheim: Beltz, 45–58.

Schwark, W. (1977): Praxisnahe Unterrichtsanalyse. Ravensburg: Otto Maier Verlag.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2004): Standards für die Lehrerbildung: Bildungswissenschaften. Zugriff am 21. 10. 2012 unter http://www.kmk. org/bildung-schule/allgemeine-bildung/lehrer/lehrerbildung.html

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2006): Bildungsstandards im Fach Deutsch für den Primarbereich (Jahrgangsstufe 4). Beschluss der Kultusministerkonferenz vom 15. 10. 2004. Zugriff am 10. 12. 2013 unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/ 2004_10_15-Bildungsstandards-Deutsch-Primar.pdf

Sinatra, G. M. / Mason, L. (2008): Beyond knowledge: Learner characteristics influencing conceptual change. In: S.Vosniadou (Hrsg.), International handbook of research on conceptual change. New York, London: Routledge, 560–582.

Tresselt, P. (2013): Schule und Recht. Zugriff am 28. 12. 2013 unter http:// www.tresselt.de

Vygotski, L. (1978): Interaction between learning and development. In: L. Vygotsky (Hrsg), Mind and Society. Cambridge, MA: Harvard University Press, 79–91.

Wason, P. C. (1960): On the failure to eliminate hypotheses in a conceptual task. In: Quarterly Journal of Experimental Psychology, Heft 12, 129–140.

Weinert, F. E. (2002): Vergleichende Leistungsmessungen in Schulen – eine umstrittene Selbstverständlichkeit. In: F. E. Weinert (Hrsg.), Leistungsmessungen in Schulen. Weinheim: Beltz, 17–32.

Wengert, H. G. (2000): Leistungsbeurteilung in der Schule. In: Bovet, G. / Huwendiek, V. (Hrsg.), Leitfaden Schulpraxis. Berlin: Cornelsen, 240–263.

Wikipedia (2014): Ziegenproblem. Zugriff am 10. 06. 2014 unter http://de. wikipedia.org/wiki/Ziegenproblem

Winter, F. (2004): Leistungsbewertung. Eine neue Lernkultur braucht einen anderen Umgang mit Schülerleistungen. Baltmannsweiler: Schneider Verlag Hohengehren.

Ziegenspeck, J. (1999): Handbuch Zensur und Zeugnis in der Schule. Bad Heilbrunn: Klinkhardt.

Zielinski, W. (1975): Die Beurteilung von Schülerleistungen. In: F. E. Weinert / C. F. Graumann / H. Heckhausen / M. Hofer (Hrsg.), Pädagogische Psychologie (Bd. 2). Frankfurt/Main: Funk-Kolleg, 877–900.