Kurt Vogelsberger

Beurteilung – mikroskopisch oder makroskopisch?

Einige grundsätzliche Gedanken zur Beurteilungsproblematik in der Lehrerinnen- und Lehrerausbildung

(reduzierte und veränderte Fassung eines Aufsatzes, veröffentlicht in: Seminar – Lehrerbildung und Schule. Hrsg. BAK. Heft 1/1996)

1. Beurteilung — ein Stiefkind allerorten?

Nicht von ungefähr sind namhafte und bedeutsame Veröffentlichungen zur Beurteilungsproblematik überwiegend älteren Datums, neuere Publikationen nehmen auf – mit Verlaub – jene „alten Hüte" immer wieder Bezug, man hat eher das Gefühl, als drehe sich das Beurteilungskarussell im Kreis, denn dass neue Gedanken und Ideen den Betroffenen Hilfe für ihre tägliche Beurteilungspraxis bieten würden. Generationen von Lehramtsanwärterinnen und -anwärtern beschäftigen sich, meist in Form von Referaten, immer wieder mit Gütekriterien, Standardabweichung, Normen und Messqualitäten, und – wenn sie zu höheren Sphären vorstoßen – auch mit Signifikanzen und Korrelationen. Skepsis ist angebracht, wenn man der Frage nachgeht, inwieweit den jungen Kolleginnen und Kollegen die Beschäftigung mit jenen Theorien konkrete Hilfen und Handlungsanleitungen für die tägliche Schulpraxis angedeihen lassen. Vor diesem Hintergrund ist es nur verständlich, dass die aus der Praxis für die Praxis „selbst gestrickten" (Überlebens-)Strategien, denen die Lehramtsanwärterinnen und -anwärter vor Ort in den Schulen begegnen, von ihnen eher angenommen werden, und dass das an der Universität oder im Studienseminar Gelernte recht schnell und mitunter zu schnell relativiert wird.
Zwischen Theorie und Praxis der Beurteilung klafft mehr als in anderen Bereichen eine enorme Lücke, was allerdings weniger ein Defizit der Schulpraxis denn ein Mangel der Theorie zu sein scheint, die es bisher nicht verstanden hat, ihre Konstrukte hinreichend transparent auf den Horizont der „Anwender" zu projizieren und praktikable Handlungsmuster und Konzepte bereitzustellen.

Dialog tut Not!

Auch die Seminardidaktik muss im Sinne eines praxisorientierten Zugangs und einer zwar theoriebegleiteten, aber primär handlungsorientierten Erschließung Wege finden, einschlägige Kompetenzen und praktikable Strategien zu entfalten. Das Heft „Seminar" 4/95 zum Thema „Seminardidaktik" des BAK bietet hier sicherlich viele Anregungen, wiewohl gerade im Zusammenhang mit „Beurteilungskompetenz" noch Wege zu ebnen sind.

Die Diskussion um Beurteilung scheint neuerdings vor dem Hintergrund des erweiterten Lernbegriffs wieder zögerlich in Gang zu kommen. Ein veränderter Lernbegriff hat zwangsläufig einen veränderten Leistungsbegriff zur Folge und erfordert ein Nachdenken, inwieweit tradierte Bewertungskonzepte modifiziert oder gar revolutioniert werden müssen. Und auch die Studienseminare kommen nicht daran vorbei, ihre Bewertungs- und Beurteilungsmechanismen in Anbetracht einer veränderten Lehrerrolle zu überdenken. Allerdings scheinen sowohl die Rahmenbedingungen als auch die Praxis selbst eher träge; wie lange wird es dauern, bis z. B. rechtliche Vorgaben der Schulordnungen zur Leistungsfeststellung und Leistungsbeurteilung geändert sind und neuen Erfordernissen Rechnung tragen?
Auch die Definition von Bildungsstandards für die schulische Bildung als Reaktion auf die Ergebnisse nationaler und internationaler Vergleichsstudien sowie curricularer Standards in der Lehrerinnen- und Lehrerausbildung macht es notwendig, Strategien der Beurteilung mit Fokus auf jene Vorgaben neu zu überdenken.

Aber wie ist es um die Beurteilungskompetenz der Ausbilderinnen und Ausbilder bestellt?
In den Studienseminaren führt die Auseinandersetzung mit der Beurteilungskompetenz der Ausbilderinnen und Ausbilder ein eher stiefmütterliches Dasein, bleibt oftmals autodidaktischen Bemühungen anheim gestellt. Sie wird in der Regel Kraft Amtes „erworben", und die in der Funktion neuen Fachleiterinnen und Fachleiter tun sich verständlicherweise anfangs recht schwer in der Bewertung von Lehrproben, Hausarbeiten und mündlichen Prüfungen und in der Abfassung von Beurteilungen der Lehramtsanwärterinnen und -anwärter. Sie orientieren sich an ihren eigenen Erfahrungen (eine mündliche Prüfung durchführen und bewerten lernt man dadurch, dass man selbst welche durchlaufen und am eigenen Leibe erfahren hat!), an vielerorts vorhandenen „Kriterienkatalogen" und an einschlägigen Mustervorlagen.
Erst die wachsende Routine und ein zunehmender Erfahrungshintergrund schaffen mehr Sicherheit. Die kritische Frage muss erlaubt sein, ob dies richtige und hinreichende Prämissen sind, um die (zumindest für die Beurteilten) so bedeutsame Beurteilungsfunktion auszuüben und auszufüllen.
Die Aus- und Fortbildung der Ausbilderinnen und Ausbilder in diesem Bereich bleibt oftmals der Initiative einzelner oder seminarinternen Bemühungen überlassen. Für die Realschulseminare des Landes Rheinland-Pfalz fand im Sommer 1995 eine dreitägige, von den Seminaren selbst vorbereitete und gestaltete Fortbildungsveranstaltung zu dieser Thematik statt, wobei die Durchführung mündlicher Prüfungen und deren Bewertung im Mittelpunkt standen. Auch diese Tagung zeigte, wie notwendig eine solche Auseinandersetzung mit Fragen der Beurteilung ist und wie dringlich sie seitens der Fachleiterinnen und Fachleiter erachtet wird.

Im Folgenden versuche ich, einige Perspektiven für die Bewertung des so überaus komplexen und vielschichtig determinierten Merkmals „Lehrerleistung" zu beleuchten, die m. E. helfen können, in den verschiedenen Beurteilungsbereichen (Beurteilungen, Hausarbeitsgutachten, Lehrproben, mündliche Prüfungen) sicherer zu werden.

2. Der Wahrheit auf der Spur?

Wenn Günter Sämmer in seinem lesenswerten Aufsatz „Handlungsorientierte Seminardidaktik“[1] von einer Unterrichtssimulation und deren Bewertung berichtet, dass trotz Nennung ähnlicher „Mängel“ (in Kurzgutachten) die Bewertungen der Stunde von „zwei minus“ bis „mangelhaft“ streuten, so deckt sich dies durchaus mit unseren Erfahrungen in vergleichbaren Veranstaltungen.

Aber: Noch so ausgefeilte Kriterienkataloge oder die Erforschung der „heimlichen Hauptkriterien“, die „implizit im Hintergrund mitspielen“, werden hieran etwas ändern.

Dass es weiterer Taten (Fortbildung etc.) bedarf, diesen eklatanten Dissens aufzulösen, mag eine zutreffende Schlussfolgerung sein, soweit es darum geht, Konsens hinsichtlich dessen herbeizuführen, was „guten Unterricht“ ausmacht. Die Zielvorstellung, dass alle Beurteiler denselben Notenvorschlag machen oder zumindest nicht gar so weit auseinander liegen, die „Treffer“ auf der Zielscheibe „Note“ sich also enger gruppieren, ist jedoch ein Irrweg. Eine solche Erwartung geht nämlich von der Prämisse aus, dass es für eine Unterrichtsstunde eine objektiv „richtige“ Bewertung, eine „Wahrheit“ gibt, die es möglichst genau zu treffen, herauszufinden gilt – und diese Prämisse ist schlichtweg **falsch**.

Nebenbei bemerkt: *Selbst wenn z. B. von fünf Beurteilern vier deckungsgleich für dieselbe Note votieren, und einer davon abweicht – woher wollen wir wissen, dass nicht ausgerechnet der eine richtig und die Mehrheit falsch liegt?*
Zur Quantifizierungen von Leistungen führt das Mehrheitsprinzip nicht unbedingt zu richtigen Entscheidungen; ebenso wenig ist die Mittelwertbildung ein a priori zuverlässiges Verfahren).

Diese Vorstellung wird von den Messtheoretikern seit Jahrzehnten unter dem Etikett „Objektivität“ verkauft (= verschiedene Beurteiler müssen zu demselben Urteil gelangen), und sie hat bei den Praktikern vor Ort ein permanent schlechtes Gewissen produziert (wenn z. B. Deutschaufsätze – und nicht nur diese – von verschiedenen Lehrerinnen und Lehrern gänzlich unterschiedlich bewertet werden) und den Kritikern schlagkräftige Argumente für das geliefert, was sie schon immer wussten: dass nämlich die Beurteilenden vielfach inkompetent seien und sich eine abenteuerliche Vielfalt von Bewertungsvorstellungen „zurechtgebastelt“ hätten.
Richtig ist im Gegenteil, dass unterschiedliche Bewertungen bei sehr komplexen und vielfach determinierten Leistungsmerkmalen (und „eine Unterrichtsstunde halten“ ist eben eine immens vielschichtige Struktur mit aberhunderten Fakten und Aspekten) normal und auch legitim sind. Subjektive Wahrnehmungen und Einschätzungen, der persönliche Erfahrungshintergrund der Beurteiler, was ihnen wichtig oder weniger bedeutsam erscheint, die vergleichende Relation zu anderen Stunden, der gemeinsame Hintergrund der Ausbildungsgenese (der z. B. „Außenbeurteilern“ wie Prüfungsvorsitzenden gänzlich fehlt), die pädagogische Verantwortung und Freiheit eines jeden Beurteilenden, legitime Beurteilungsspielräume u. v. a. m. – all das sind Fakten, die zwangsläufig zu unterschiedlichen Einschätzungen führen müssen, das Gegenteil wäre verwunderlich.

Es kann also im Sinne einer falsch verstandenen Objektivität nicht Ziel sein, alle Beurteiler einer Lehrprobe zu einem einhelligen Urteil gelangen zu lassen. Solche Bemühungen wären gewiss von vornherein zum Scheitern verurteilt.

Eine Wahrheit gibt es nicht. Für die **Beurteilung und Bewertung von beobachtetem Unterricht** (in der Regel wird die Note von einer Kommission erteilt) scheinen mir demgegenüber folgende Aspekte von Bedeutung:

- „Objektivität“ wird dadurch erreicht, dass mehrere subjektive Urteile zum Tragen kommen und zu einem möglichst weit gehenden Konsens geführt werden (dazu müssen Prüfungsvorsitzende, Seminarleiterinnen und -leiter, ... bereit sein, sich einzulassen auf das, was andere meinen).
- Die Beurteiler müssen offen sein für abweichende Sichtweisen und andere Meinungen gelten lassen (dass Beurteiler abfällig an der Kompetenz anderer zweifeln, wenn diese abweichend votieren und argumentieren, ist der Sache nicht dienlich).
- Die Beurteiler müssen den ständigen Dialog suchen (mit Anwärterinnen und Anwärtern, mit Fachleiterinnen und Fachleitern) und versuchen, möglichst breiten Konsens zu finden darüber, was guten Unterricht ausmacht (beurteilungsrelevant ist nur das, was auch vorher, z. B. im Verlauf der Ausbildung, als bedeutsam galt; es werden nicht unvermittelt neue Kriterien aktiviert).
- Die Maßstäbe für Qualität haben sich in erster Linie an der Sache und an der Lerngruppe zu orientieren (eine didaktische oder methodische Entscheidung ist z. B. nicht allein deswegen gut oder schlecht, weil sie den Vorstellungen der Fachleiterin bzw. des Fachleiters entspricht oder zuwiderläuft, weil sie „gefällt“ oder „nicht gefällt“).
- Sachfremde Erwägungen müssen außen vor bleiben, nur die Stunde selbst steht zur Disposition (dazu zählen z. B. Voreinschätzungen, die eine wesentliche Fehlerquelle darstellen: „ ... sonst waren die Leistungen wesentlich besser/schlechter, also ...“; dazu zählen auch taktische Überlegungen wie: „ ... wenn wir jetzt ... geben, dann ...“). Kritiker unserer Bewertungspraxis monieren, mitunter durchaus mit Recht, dass Lehrproben und mündliche Prüfungen oft nur dazu dienen, das zu bestätigen, was die Prüferinnen und Prüfer vorher schon wussten.
- Das Taktieren mit Notenvorschlägen (z. B. votieren abweichend von der eigenen Überzeugung, um in einer Kommission letztlich dort zu landen, wo man wollte) muss unterbleiben: alle Beurteiler vertreten offen und ehrlich ihre Auffassung.

[1] Seminar 4/95, Seite 36 ff.

- Eigene Befindlichkeiten müssen ebenso zurückstehen wie Verhaltensweisen, die der Beziehungsebene innerhalb der Gruppe der Beurteiler entspringen (z. B. man hält sich zurück, weil ...; man versucht sich durchzusetzen, weil ...).

... und vor allem:

- „Lehrproben" müssen „normaler" Unterricht sein (z. B. es wird nicht „vorgebaut", es wird kein besonderer Aufwand getrieben, es muss kein „gutes" Thema sein, die Lerngruppe wird nicht „vergattert", ...), nur dann erlauben sie zuverlässige Rückschlüsse auf die Qualifikation der Kandidatinnen und Kandidaten.

Die vorgenannten Gesichtspunkte mögen manchen Leser überraschen oder gar befremden. Sie spielen nach meiner Erfahrung (in inzwischen einer Unzahl von bewerteten Stunden) eine größere Rolle und haben mit „Objektivität" und der Wahrheit mehr zu tun, als man (und manchmal auch ich) eingestehen will.

3. Mikro- oder makroskopische Beurteilung?

In allen Beurteilungsfeldern (Beurteilung der Anwärterinnen und Anwärter, schriftliche Hausarbeiten, Lehrproben, mündliche Prüfungen) sind in der Praxis der Beurteiler vergleichbare Strukturen resp. Techniken bei der Notenfindung zu beobachten. Am Beispiel der Bewertung von Lehrproben möchte ich die Extreme, die von einer sehr detaillierten Aufschlüsselung einzelner Aspekte (Mikroskopie) bis hin zu sehr globalen und pauschalen Feststellungen (Makroskopie) reichen, im Folgenden beschreiben.

Die letztgenannte, makroskopische Technik ist sehr schnell charakterisiert:
Beurteiler sind mit ihrer Note recht schnell „bei der Hand", sagen (oft im Brustton tiefster Überzeugung), „das ist ..." und liefern dazu im Nachsatz noch ein oder zwei Argumente, wohl mehr, um ihre Auffassung zu bekräftigen denn sie stichhaltig zu begründen. Die genannten Argumente sind dabei zudem nicht selten von recht nachgeordneter Bedeutung.
Dass diese wohl eher spontane Technik der „Notenfindung" einer genaueren Überprüfung hinsichtlich Objektivität nur selten standhält, liegt auf der Hand.
Leider ist es so, dass diese Beurteiler sich oft genug auch noch durchsetzen, machen sie doch einen sehr kompetenten Eindruck und halten, im Gegensatz zu anderen, die sich unsicher fühlen und noch keine Entscheidung getroffen haben, den Dialog brauchen, mit ihrer Meinung und ihrem Anspruch, „richtig" zu liegen, nicht hinter dem Berg.

Die „mikroskopische" Bewertungstechnik versucht im gewissenhaften Streben nach Vollständigkeit möglichst viele Aspekte, womöglich alle, zu berücksichtigen. Oftmals werden hierzu die wohl in jedem Seminar vorhandenen „Kriterienkataloge zur Beobachtung und Bewertung von Unterricht" herangezogen, die atomistisch viele, oft fast unendlich viele Kriterien auflisten. Ein solcher Kriterienkatalog existierte auch an meinem ehemaligen Studienseminar und leistete durchaus seine Dienste, ein „kleiner Auszug" ist nachstehend dargestellt:

Beobachtung und Beurteilung von Einzelstunden

1. Unterrichtsplanung

> die Stunde ist fachlich und stofflich gut vorbereitet
> die Lernziele sind stichhaltig begründet
> die Stunde ist richtig in den größeren Zusammenhang eingeordnet
> die didaktische Reduktion auf die Ebene der Schüler ist ohne fachliche Fehler oder einseitige Schlussfolgerungen gelungen
> die Lernziele sind in Bezug auf Inhalt und Niveau angemessen
> die Voraussetzungen beim Schüler (stofflich, methodisch) sind hinreichend beachtet
> die Stunde ist methodisch sinnvoll und folgerichtig strukturiert (Artikulation)
> der Einstieg ist motivierend und problembezogen gewählt
> Wiederholung, Übung und Anwendung sind vorgesehen
> ein Wechsel der Aktions- und Sozialformen ist vorgesehen
> Lernzielkontrollen sind vorgesehen
> Experimente, Quellen, Anschauungsmittel, ...werden angemessen herangezogen
> schwierige Stellen im Lernprozess werden im Vorhinein erkannt
> Aspekte der Differenzierung werden berücksichtigt
> es werden Maßnahmen ergriffen, die geeignet sind, die Lehrerzentrierung zu reduzieren

2. Durchführung

2.1. Lehrerverhalten

> Kontakt zu den Lernenden ist vorhanden, der Lehrton ist bestimmt, höflich und partnerschaftlich
> der Überblick über das Geschehen ist vorhanden, die Lehrkraft ist nicht nur im Lehrstoff befangen
> Arbeitsdisziplin wird hergestellt, Erziehungsmaßnahmen werden angemessen gehandhabt
> unvermutete Schwierigkeiten, situative Probleme, weltergehende Fragen werden bewältigt
> der Sprechanteil der Lehrkraft ist angemessen gering, jener der Schülerinnen und Schüler hoch
> die Lehrersprache ist korrekt in Form und Inhalt, anschaulich und akzentuiert
> Mimik, Gestik und Sprache sind abwechslungsreich und ausdruckskräftig
> die Lehrkraft nimmt sich Zeit, vermeidet Hektik und stoffliche Überfrachtung

2.2. Beherrschung der Unterrichtstechniken

> der Einstieg motiviert und führt zügig zur Problemstellung
> die Schülerinnen und Schüler sind rechtzeitig über die Zielsetzung informiert
> Medien, Experimente, ... werden im rechten Umfang, zur rechten Zeit ökonomisch eingesetzt
> das Tafelbild ist übersichtlich, repräsentativ, logisch und stellt eine Lernhilfe dar
> offene Fragen und Impulse werden geschickt formuliert und wirksam eingesetzt
> die Lernenden erhalten eindeutige, differenzierte und abwechslungsreiche Rückmeldungen
> die Abfolge der Lernschritte und die Überleitungen erscheinen für die Schülerinnen und Schüler logisch zwingend und sind motiviert
> die Mehrzahl der Schülerinnen und Schüler wird zur aktiven Mitarbeit gebracht
> Ergebnisse werden angemessen und rechtzeitig gesichert (Hervorhebung, Wiederholung, Tafel, ...)
> Wiederholungen, Übungen, Anwendungen werden im notwendigen Umfang durchgeführt
> Lernziele werden überprüft, die Ergebnisse der Überprüfung bei der Weiterführung des Unterrichts berücksichtigt
> Immanente Wiederholung, Reorganisation des Vorwissens und Transfer finden statt
> die Bedingungen des Begriffs- und Regellernens werden beachtet
> Hausaufgaben werden kontrolliert, sinnvoll und rechtzeitig gestellt, eindeutig formuliert
> Schülerleistungen werden gewürdigt und treffend bewertet
> der Unterricht ist lebensnah, aktualitätsbezogen, ideenreich, ermöglicht Realbegegnungen
> die Sprache der Schülerinnen und Schüler (Muttersprache/Fachsprache) wird durchgängig gefördert
> das Anspruchsniveau ist hoch, Lernende werden gefordert, müssen etwas leisten
> die Kommunikation wird gefördert (deutlich/laut sprechen, zuhören, aufeinander reagieren)
> die Lehrkraft tritt wo immer möglich in den Hintergrund, gibt Freiräume, lässt Schülerinnen und Schüler agieren

3. Besprechung/Nachbereitung

> Stärken und Schwächen der Planung werden erkannt
> Stärken und Schwächen der Durchführung werden erkannt
> der Lernerfolg wird treffend beurteilt
> Verbesserungsvorschläge/Alternativen werden eingebracht und begründet
> Anregungen werden aufgegriffen und verarbeitet
> Lernpsychologische/soziologische Gesichtspunkte (Schüler, Lehrer-Schüler-Verhältnis) werden sachgerecht dargestellt
> Allgemeine fachdidaktische und -methodische Gegebenheiten werden angemessen erörtert und zu Einzelentscheidungen in Beziehung gesetzt

Falls Sie das nicht oder nur mit der Lupe lesen können, so ist das intendiert; die Darstellung veranschaulicht die komplexe und vielfältige Struktur. Jeden Teilaspekt nach seiner „Erfüllungsqualität" zu bewerten, womöglich mit Punkten, und diese dann zu einer „Gesamtbewertung" aufzuaddieren, wird wohl (hoffentlich) niemand in den Sinn kommen?
Es spielt im Grunde keine Rolle, welche Kriterienkataloge verwendet werden, ob eigene individuelle, ob in einem größeren Bereich abgestimmt, ob konsequent an Kriterien guten Unterrichts orientiert (z. B. nach Hilbert Meyer oder nach von Instanzen der Qualitätssicherung erstellten), das Grundprobleme der „unüberschaubaren" Vielfalt ist allen Katalogen zu eigen (denn die Autoren derselben wollen ja keinesfalls „wichtige" Aspekte außen vor lassen.

Und dennoch – die Anhänger der Mikroskoptechnik praktizieren in mehr oder minder ähnlichen Varianten (wobei oft eigene „Privat-Kataloge" und „Beurteilungsraster" zum Einsatz kommen) genau dieses Verfahren, um zu ihrer Stundenbewertung zu finden. Und sie liegen, folgte man den Erkenntnissen der Testtheoretiker, noch nicht einmal so falsch: je kleiner die „Test-Items", desto eher kann davon ausgegangen werden, dass verschiedene Beurteiler zu vergleichbaren Urteilen gelangen (allerdings sind noch so scheinbar kleine Items wie z. B. *„das Tafelbild stellt eine Lernhilfe dar"* noch überaus komplex und lassen immer noch viele Interpretationen zu).

Auch diese Verfahrensweise zur Ermittlung einer Note wird, auch wenn sie auf den ersten Blick den Anschein von mehr Objektivität erweckt, der Sache nicht gerecht. Um mit Aristoteles zu reden: Das Ganze kann eben mehr (oder auch weniger) sein als die bloße Summe und Aneinanderreihung einzelner Teile; gänzlich unbeantwortet bleiben z. B. bei solch atomistischen Zerlegungen einer Stunde die Fragen, inwieweit die einzelnen Aspekte gleichrangig sind oder unterschiedlich gewichtet werden müssten, oder inwieweit wichtige „Hauptkriterien" (z. B. Lernklima, Verständlichkeit in der Wissensvermittlung, ...) unterschiedlich repräsentiert sind. Und was ist, wenn viele Aspekte einer solchen Liste positiv zu bewerten sind, einige wenige Fehler aber wie eine Seuche übermächtig werden und die Stunde missraten lassen?
Im Übrigen ist ein derart komplexer Mechanismus schlicht und einfach „unhandlich". Ich für meinen Teil kann und konnte mit solchen Kriterienkatalogen nicht umgehen, jedenfalls nicht, um zu einer treffenden Einschätzung einer Unterrichtsstunde zu gelangen.

Der rechte Weg zwischen makroskopischer und mikroskopischer Bewertungsstrategie liegt, wie so oft, wohl in der Mitte. Es ist also eine „Bewertungsstrategie mittlerer Komplexität" anzustreben.
Einige wenige globale „Hauptkriterien" müssen definiert werden. Sich darauf zu einigen, welche das sein könnten oder müssten, dürfte nicht allzu schwer sein (z. B. Lernklima, Darstellungskraft, Verständlichkeit in der Wissensvermittlung, Anforderungsniveau, Schülerzentrierung, Strukturierung des Lernprozesses, o. Ä.). Um während der Stunde zu diesen Hauptkriterien Beobachtungen, Eindrücke, Unterpunkte zu sammeln, bietet sich die Methode des „Mindmapping" an (vgl. hierzu den Aufsatz von Hartmut Fischer im Heft 1/1996 des BAK).
Zu jedem Hauptkriterium ist sodann eine summarische Einschätzung vorzunehmen, und zwar mit möglichst wenigen Kategorien in numerisch angenäherter Form (z. B. dreistufig mit **+/ 0 / –**).
Diese lassen sich dann relativ leicht in eine Note transformieren, wobei noch die Option bleiben sollte, unterschiedliche Gewichtungen einzubringen (wenn z. B. die Tatsache, dass die Schülerinnen und Schüler gänzlich unterfordert waren, den Wert der Stunde insgesamt nachhaltig in Frage stellt, o. Ä.).

4. Die „Strategie mittlerer Komplexität" bei mündlichen Prüfungen

Die „Bewertungsstrategie mittlerer Komplexität" lässt sich ebenso bei allen anderen Prüfungsteilen wie personenbezogenen Beurteilungen, Hausarbeiten und mündlichen Prüfungen anwenden. Im Folgenden möchte ich dies nochmals am Beispiel der mündlichen Prüfungen verifizieren. Dabei will ich die Frage nach der Vorbereitung und der inhaltlichen wie methodischen Gestaltung mündlicher Prüfungen außen vor lassen, obwohl sie eng verwoben ist mit der Bewertungsproblematik.

Die oben dargelegten Techniken der Notenfindung werden gleichermaßen bei mündlichen Prüfungen praktiziert.
Da gibt es Prüferinnen und Prüfer, die am Ende spontan und ohne differenzierte Betrachtung **wissen**, welche Note das war. Sofern im Protokoll zur mündlichen Prüfung detaillierte Bewertungsvermerke anzubringen sind (wie z. B. „b" = beantwortet, „m.H.b." = mit Hilfen beantwortet, „s.b." = umfassend selbstständig beantwortet, ..., oder Ähnliches), werden diese im Nachhinein entsprechend und passend vermerkt oder korrigiert.
Diese Spezies „Beurteiler" wird vermutlich auch bei der Lektüre dieses Beitrags recht schnell zu dem Schluss kommen, dass er überflüssig sei.
Da gibt es aber auch Prüferinnen und Prüfer, die (anhand des Protokolls) in akribischer Weise den Prüfungsverlauf und die Äußerungen des Prüflings zu erinnern suchen und sich Schritt für Schritt an eine Note „herantasten". Diese Technik lässt wohl eher als die vorgenannte erwarten, dass eine einigermaßen treffende und „richtige" Note gefunden wird. Allerdings ist es gerade bei mündlichen Prüfung immens schwierig, sich in der Vielzahl dessen, was geäußert wurde, zurechtzufinden (es gibt 30-minütige mündliche Prüfungen, in denen ca. 3.000 bis 4.000 Wörter gesprochen werden, das sind 6 bis 8 eng beschriebene Schreibmaschinenseiten!).

Dieses relative „Chaos" braucht eine Struktur, will man es einer einigermaßen objektiven Bewertung zuführen.
Da in wohl jeder mündlichen Prüfung mehrere verschiedene Fragen- oder Themenkomplexe angegangen werden, bietet es sich an, an dieser Struktur auch die mittlere Bewertungsebene zu orientieren, d. h.

- in einem ersten Schritt wird jeder bearbeitete Themenbereich in einer (nicht zu stark differenzierten) Bewertung quantifiziert,

- in einem zweiten Schritt werden sodann die Teilbewertungen, eventuell noch unterschiedlich gewichtet, zu einer Note „aufaddiert".

Im **ersten Schritt** werden zu einzelnen Themenkomplexen oder Unterpunkten Bewertungsvermerke als **subjektive Einschätzungen** ermittelt (welche Kategorien man hier verwendet, ob „s.b./b./m.H.b./t.b./n.b." oder „0/1/2/3 ... Punkte" oder „+/0/–", oder ob man nach „lückenhaften / hinreichenden / guten Kenntnissen und defizitärer / hinreichender / vielschichtiger Anwendung" differenziert, ist letztlich unerheblich und bleibt der Konvention überlassen).
Dieser Vorgang ist zugegebenermaßen mit vielen Unwägbarkeiten verbunden, müssen doch Reduktionen, Glättungen, Schwerpunktsetzungen etc. vorgenommen und Schwierigkeitsgrade, Hilfegrade, Fehler etc. eingeschätzt werden. Daran wird sich jedoch nichts ändern lassen, und es ist sogar mit den Grundsätzen „objektivierter Leistungsmessung" als „fachlich-pädagogische Wertung" vollends vereinbar.
Wichtig erscheint mir, dass in diesem ersten Bewertungsschritt möglichst wenige Kategorien verwendet werden und nicht zu fein differenziert wird. Auch wenn Unterschiede zwischen den Probanden damit zunächst nivelliert werden, wächst die Bewertungssicherheit doch beträchtlich. Und die Differenzierung stellt sich wieder ein, wenn das Mosaik einzelner Bausteine sich zu einem Gesamtbild fügt (siehe „zweiter Schritt").
(Ein kleiner Exkurs am Rande: Die in vielen Bundesländern übliche Bewertung in einer 16– (0 bis 15 Punkte) oder x-stufigen Punkteskala führt in der Praxis zu recht abenteuerlichen „Zielübungen". Ich für meinen Teil sehe mich außer Stande, die Äußerungen eines Prüflings zu einem Fragenkomplex (oder eine Lehrprobe, siehe Punkt 4.) in 16 verschiedene Qualitätsstufen einzuordnen, und ich nehme an – zumindest hoffe ich das –, dass es vielen anderen auch so geht. Wenige Kategorien müssen her (drei bis fünf?, nur zwei wären mir am liebsten), das traue ich mir mit einiger Bewertungssicherheit zu! Die 16 Stufen ergeben sich dann schlicht und einfach und ebenso treffsicher durch „Bündelung", Zusammenfassung, Addition. Projiziert man dieses Procedere auf vertraute Mechanismen wie Klassenarbeiten: kaum eine Fachlehrkraft käme auf die Idee, eine einzelne Aufgabe mit 0 bis 15 Punkten zu bewerten; diese Feinstufung ergibt sich erst durch Aufsummierung aller Aufgaben.)

Der **zweite Schritt** allerdings, die Bewertungsvermerke zu einer Note aufzurechnen, darf (so auch der Tenor einschlägiger Verwaltungsgerichtsurteile), bis auf Gewichtungen, nicht nochmals mit einer zweiten fachlich-pädagogischen Wertung einhergehen, er muss im Wesentlichen „numerisch" vollzogen werden. Dazu ist es allerdings erforderlich, die Bewertungsvermerke zu quantifizieren (z. B. in Punkte). Was heißt also z. B. „mit Hilfen beantwortet", was ist das wert? — im Unterschied zu z. B. „teilweise beantwortet"?

Um das noch mal an einem Beispiel deutlich zu machen: Sie müssten in der Lage sein, die folgenden Bewertungsschemata (geprüft wurden vier Themenkomplexe)

1. Orientierungsstufe	*b.*
2. Erweiterter Lernbegriff	*m.H.b.*
3. Veränderte Jugend heute	*s.b.*
4. Versetzungsentscheidung	*m.H.t.b.*

Kurzfassung eines Protokolls, Beispiel 1:
Die vier Themenkomplexe wurden global bewertet

1. Aufgaben und Funktionen der Orientierungsstufe	
• Karikatur; Problembereiche benennen	*m.H.b.*
• Schlüsselbegriffe nennen (Aufgaben)	*m.H.t.b.*
• Praktische Umsetzung	*m.H.t.b.*
2. Erweiterter Lernbegriff	
• Neuer Lernbegriff – alter Lernbegriff ?	*m.H.b.*
• Erweiterten Lernbegriff definieren	*m.H.t.b.*
3. Veränderte Jugend heute	
• Charakterisierung „Jugend heute"	*b.*
• Schülerinnen und Schüler typisieren (5.)	*m.H.b.*
4. Versetzungsentscheidung	
• Fallbeispiel entscheiden	*b.*
• Versetzung in besonderen Fällen	*n.b.*

Kurzfassung eines Protokolls, Beispiel 2:
Die vier Themenkomplexe wurden detailliert bewertet

in eine Note zu „übersetzen"[2], und es wäre nicht schlecht, wenn Ihre Kolleginnen und Kollegen zu genau demselben Resultat kämen.
Wenn Sie jetzt einwenden „wie soll das möglich sein, wo ich doch gar nicht weiß, was im Einzelnen gefragt und wie geantwortet wurde", dann ist das genau der Punkt: darauf darf es – bis auf unterschiedliche Gewichtung der vier Komplexe – jetzt und bei diesem Schritt keinesfalls mehr ankommen!
Verschiedene Gewichtungen (z. B. die Einschätzung eines Teilbereiches wird doppelt gewertet) dürfen nicht intuitiv vorgenommen werden, sie müssen von der Sache her begründet sein, etwa aufgrund des zeitlichen Umfangs, der Bedeutung des Sachverhalts im Kontext der Lehrerqualifikation oder des Schwierigkeitsgrades usf.

In allen Beurteilungsbereichen ist diese Strategie mittlerer Komplexität (Splitten der komplexen Leistung in Teilsequenzen, welche einzeln in wenigen Kategorien eingeschätzt und sodann zur „Gesamtleistung" aufgerechnet werden) m. E. die einzige Möglichkeit, Ansprüchen objektivierter Leistungsmessung in Ansätzen gerecht zu werden.

[2] Meine „Lösung": Sie müssten (bei gleicher Gewichtung) in Beispiel 1 bei „befriedigend" landen, in Beispiel 2 bei „ausreichend"

5. Über das Huhn und das Ei und die Mathematik

Die philosophische Frage, was zuerst da war, ob Huhn oder Ei (sprich: die Note oder ihre Begründung), ist im Zusammenhang mit Beurteilungen und Bewertungen eine durchaus entscheidende. Die spontane und intuitive und von Voreinschätzungen gefärbte Gewinnung einer Lehrproben-Note, die im Nachhinein begründet wird, ist sicherlich kein guter Weg, zu gerechten und vergleichbaren Bewertungen zu gelangen (denn: auf der Suche nach Begründungen und zementierenden Argumenten zu einer vorgefassten Einschätzung werden wir in solch komplexen Leistungen immer fündig; in der Sprache der Anwärterinnen und Anwärter: „ ... die finden immer was auszusetzen ... " – wie wahr!). Es mag weit hergeholt klingen, aber auch der Tenor „der Fachleiter oder die Fachleiterin schlägt eine Note vor und begründet diese ..." leistet dem Vorschub.
Umgekehrt wäre besser: aus der Einschätzung von Teilaspekten und der Summe dieser leiten wir die Note ab.
Für die Fertigung einer personenbezogenen Beurteilung über eine Anwärterin oder einen Anwärter ist es mehr als nur eine Frage des Arbeitsablaufs, ob man mit der Note beginnt und diese dann verbal ausformt, oder ob man zu zentralen Bereichen der Qualifikation Aussagen trifft und aus der „Summe" der einzelnen Elemente die treffende Note ableitet.
Mit einer mathematischen Metapher lässt sich dieser Sachverhalt „Huhn oder Ei" sehr schön darstellen: Zu einem Term (eine Anordnung von Zahlen und Rechenzeichen) gehört in der Regel nur ein einziger, ganz bestimmter Wert; geht man jedoch von dem Zahlenwert aus, so kann man unendlich viele verschiedene Terme finden, die diesen Wert besitzen.

Es gibt gewiss viele Beurteiler, die die Nase rümpfen, wenn man ihnen mit „Mathematik" kommt, denn „ ... erstens ist das viel zu kompliziert und zweitens kann man doch Lehrerleistung nicht mathematisch abbilden oder wie auf einem Kassenzettel aufrechnen ... ".
Andererseits zollen wir mit Notenberechnungen *auf zwei Dezimalen genau* gerade eben diesem Götzen Tribut, und zwar in einer Art und Weise, die die Grenze zur Perversion und zum mathematischen Nonsens bereits weit überschritten hat (wir „schätzen" auf bestenfalls Meter, gleichsam Pi mal Daumen, und berechnen dann daraus auf Zentimeter genau Durchschnittswerte, und wir machen, um es auf die Spitze zu treiben, von der zweiten Nachkommastelle auch noch Einstellungs- und Lebenschancen abhängig. Nebenbei bemerkt: Es wird höchste Zeit, von solchem Tun wider besseres Wissen abzulassen und zu anderen Mechanismen zu finden).

Es geht nichts daran vorbei: Beurteilung und Bewertung heißt abschätzen, abwägen, vergleichen, quantifizieren, messen etc., und dies bedeutet, dass man ohne eine kleine Portion Mathematik nicht auskommt – aber bitte in Maßen und denklogisch und sachgerecht. Wo die Mathematik ihren unverzichtbaren Platz hat? Ich hoffe, es ist mir gelungen, dies deutlich zu machen.

6. Anhang: Einige Quantifizierungen

Im Folgenden sollen noch einige mögliche Quantifizierungen für mündliche Prüfungen vorgestellt werden. Dabei möge man bedenken, dass es hierfür im Grunde viele Alternativen gibt. Was „gerecht" und sachgerecht ist, ist nicht eindeutig, darüber lässt sich (mit Beurteilern und Beurteilten) endlos streiten und diskutieren – muss schlicht und einfach via Konvention festgelegt werden. Wichtig ist (und das ist gerecht), dass eine vereinbarte Quantifizierung bei allen Probanden in gleicher Weise angewandt wird.

	selbstständig beantwortet	beantwortet	teilweise beantwortet	nicht beantwortet
ohne Hilfe	10-9	8-7		
mit Hilfe		6-5	4-3	
mit vielen Hilfen			2-1	0

In der Matrix aus 4 Antwortqualitäten und 3 Hilfegraden sind von vornherein 6 Felder auszuschließen (z. B. eine selbstständige Beantwortung mit Hilfen gibt es nicht; ebenso ist eine teilweise Beantwortung ohne Hilfe auszuschließen, weil die Prüferin oder der Prüfer bei unvollständiger Beantwortung versuchen wird, dem Prüfling „auf die Sprünge zu helfen"; usf.). Im obigen Modell ist in jedem möglichen Feld noch eine Differenzierungsmöglichkeit vorgesehen. Denkbar wäre auch:

	selbstständig beantwortet	beantwortet	teilweise beantwortet	nicht beantwortet
ohne Hilfe	5	4		
mit Hilfe		3	2	
mit vielen Hilfen			1	0

Eine andere Struktur zur Einschätzung der Prüfungsleistungen wäre eine Differenzierung nach Kenntnissen und Anwendung derselben, z. B.

Kenntnisse	keine 0	lückenhaft 1	hinreichend 2	gut 3	umfassend 4
Anwendung	defizitär 0	eingleisig 2	befriedigend 4	vielschichtig 6	

Die Summe der so (oder auch anders) ermittelten Punktwerte könnte mit einer Prozentskala (wiederum Konvention) in eine Note umgerechnet werden; z. B.

%	0-12	13-31	32-50	51-68	69-87	88-100
Note	6	5	4	3	2	1

Solche Quantifizierungsmodelle sollen aber keineswegs „sklavisch" angewandt werden, sie mögen Anhaltspunkte und Orientierungen bieten und die Bewertungssicherheit verbessern. Auch für mich ist es keine sonderlich attraktive Vorstellung, dass Prüferinnen und Prüfer am Ende einer mündlichen Prüfung mit dem Taschenrechner der Sache zu Leibe rücken.

Kurt Vogelsberger
Realschulrektor i. R.
ehemaliger Leiter eines Staatlichen Studienseminars
für das Lehramt an Realschulen

© Kurt Vogelsberger 2009