

— Der bisherige Kenntnisstand über die Ursachen für das Versagen bei Abschlußprüfungen erlaubt noch keine konkreten Aussagen darüber, daß die zentrale Funktion beruflicher Abschlußprüfungen in der Steuerung mittels Selektion zu sehen ist (vgl. LENNARTZ 1976). Eine normorientierte Leistungsmessung hingegen durfte nur dann im Vordergrund stehen, wenn die Bewertungsverfahren für eine Selektionsentscheidung im o. g. Sinne herangezogen werden sollten.

Aus diesen Überlegungen ergeben sich erhebliche Konsequenzen auch für die Leistungsbewertung. So wäre es beispielsweise nicht gerechtfertigt, bei beruflichen Zwischen- und Abschlußprüfungen — als ausbildungzielorientierten Leistungsbewertungen — Aufgabensätze so zu konstruieren oder bereits vorliegende Ergebnisse so zu transformieren, daß eine Normalverteilung (die vielgenannte Gauß'sche Glockenkurve) der Punkte oder Noten entsteht. Dagegen sind Verfahren von Interesse, mit denen aufgrund neuerer teststatistischer Ansätze Bewertungsschlüssel in Form einfacher Tabellen konstruiert werden können, ohne daß es erforderlich ist, zur Gewinnung eines Vergleichsmaßstabs vorher die entsprechenden Leistungen bei einer größeren Gruppe von Prüfungsteilnehmern festzustellen und auszuwerten. Bei einem solchen Bewertungsschlüssel werden die Zahl (oder der Anteil) der gelösten Aufgaben und der Grad der Zielerreichung („Erreichung“ oder „Nichterreichung“ oder Noten als unterschiedliche Grade der Zielerreichung) in Beziehung gesetzt. Beispiele finden sich bei KLAUER 1972 und HERBIG 1975, eine Übersicht über den Forschungsstand hinsichtlich von Bewertungsmodellen für die kriterienorientierte Leistungsmessung enthält das Sammelreferat von MESKAUSKAS 1976.

Während somit eine ausbildungzielorientierte Bewertung auf die Frage abzielt, ob bzw. inwieweit ein Jugendlicher — unabhängig von anderen — das jeweils vorgegebene Ausbildungziel erreicht hat, geht es bei der **normorientierten Leistungsmessung** darum, Unterschiede zwischen den Auszubildenden möglichst genau zu ermitteln. Eine solche Bewertung kann etwa bei eignungsdiagnostischen Untersuchungen Verwendung finden.

Begriffe wie Lernfortschrittskontrollen, Lernerfolgskontrollen und Lernzuwachsmessung weisen auf eine **prozeßorientierte Leistungsbewertung** hin. Wenn man berufliche Bildung als Veränderungsprozeß ernstnimmt und die Leistungsmessung

in diesem Zusammenhang sieht, müssen solche Bewertungen in den Mittelpunkt des Interesses treten. Um beispielweise Unterschiede zwischen Eingangsleistungen und Ergebnissen von Zwischen- und Abschlußprüfungen bei einem Auszubildenden zu bewerten, waren spezielle Verfahren notwendig, die noch weitgehend fehlen.

Zusammenfassend ist festzuhalten, daß — je nach den Zielen, die mit der Erfassung von Lernleistungen in der beruflichen Bildung verfolgt werden — auch unterschiedliche Bewertungsprozeduren erforderlich sind. Es würde den Rahmen dieser Anmerkungen sprengen, die zur Zeit praktizierten Verfahren (100-Punkte-System, Notensystem) auf diesem Hintergrund zu analysieren und — falls es erforderlich ist — Änderungsvorschläge zu entwickeln und auf ihre Brauchbarkeit hin zu prüfen.

#### Literatur

- Brand, Peter, 1975: Programmierte Prüfungen im Spiegel der Erfahrung und der Bewahrung. In: *Programmierte Püfungen: Problematik und Praxis. Schriften zur Berufsbildungsforschung*. Band 35. Schroedel Hannover S. 181—252
- Fricke, Rainer, 1974: Kriteriumsorientierte Leistungsmessung Kohlhammer. Stuttgart
- Herbig, Manfred, 1975: Zensierung nach dem binomialen Testmodell. In: *Lernzielorientierter Unterricht*, H. 1, S. 1—12
- Herbig, Manfred, 1976: Praxis lehrzielorientierter Tests. Schwann: Düsseldorf.
- Klauer, Karl Josef, 1972 Zur Theorie und Praxis des binomialen Modells lehrzielorientierter Tests. In: Klauer et. al.: *Lehrzielorientierte Tests* Schwann Düsseldorf, S. 161—195.
- Klauer, Karl Josef; Fricke, Rainer; Herbig, Manfred; Ruprecht, Helmut, Schott, Franz, 1972 *Lehrzielorientierte Tests* Schwann Düsseldorf.
- Lennartz, Dagmar, 1976: Zur Bestimmung von Determinanten des Prüfungsversagens — Überlegungen zu einem Untersuchungs-Ansatz. Manuskriptdruck Bundesinstitut für Berufsbildungsforschung Berlin
- Leth, August, 1975 Programmierte Prüfungen mit dem Normtest-Electronic-System. In: *Programmierte Prüfungen: Problematik und Praxis. Schriften zur Berufsbildungsforschung*, Bd. 35. Schroedel Hannover S. 255—331
- Lienert, Gustav A., 1969: Testaufbau und Testanalyse 3. Aufl. Beltz. Weinheim
- Meskauskas, John A., 1976: Evaluation Models for Criterion-Referenced Testing. Views Regarding Mastery and Standard-Setting. In: *Review of Educational Research*, Jg. 46, H. 1, S. 133—158
- Nibbrig, Bernhard, 1976: Gutachten zum Problem der programmierten schriftlichen Kaufmannsgehilfenprüfung. Köln: Institut für Berufs-, Wirtschafts- und Sozialpädagogik der Universität zu Köln
- Pawlik, Kurt (Hrsg.), 1976: *Diagnose der Diagnostik* Klett: Stuttgart
- Rüdiger, Dietrich, 1976: Prozeßdiagnose als neues Konzept der Lernfähigkeitsdiagnose. Referat auf dem 30. Kongreß der Deutschen Gesellschaft für Psychologie vom 19.—23. September 1976 in Regensburg.

Ernst Hoffmann

## Zur Frage von Punktbewertungsrichtlinien

Seit vielen Jahren sind Versuche großer Industrieunternehmen der Metall- und Elektroindustrie bekannt, die Leistungen ihrer im Bundesgebiet verstreuten Lehrwerkstätten anhand der Prüfungsergebnisse vergleichen. Die Konzernzentrale kann Ergebnisse durchaus konkret erfassen, so daß der Gedanke solcher Vergleiche naheliegt; vor allem, wenn die Zentrale davon überzeugt ist, daß für diese Gruppe von Berufen bundeseinheitliche Aufgaben und bundeseinheitliche Bewertungsrichtlinien verwendet werden. Außer Ärger erbringen aber solche Vergleiche nichts. Sie fußen schlicht auf einer Überschätzung von Prüfungsresultaten, die anders als in Technik und Sport mit Stoppuhr und Bandmaß weder exakt erfassbar noch exakt vergleichbar sind.

Oder sollte das falsch sein? Sollte es möglich sein, mit genauen und einheitlichen Bewertungsvorschriften Prüfungsergebnisse von Ort zu Ort, von Lehrberuf zu Lehrberuf, von

Lehrbetrieb zu Lehrbetrieb sowie von Termin zu Termin vergleichbar zu machen? Langjährige Erfahrungen lassen da zweifeln.

Bewertungsregelungen sind notwendiger Bestandteil des Prüfungswesens und haben vierlei Sachverhalte zum Gegenstand. Überlegungen über ihre Bestgestaltung füllen z. B. beim Deutschen Industrie- und Handelstag viele dicke Aktenordner. Bewertungsrichtlinien müssen auch aus naheliegenden Gründen für Metallberufe andere Sachverhalte regeln als für Bau-, Labor- oder reine Büroberufe. Somit sind Vergleiche von einander fremden Berufsgruppen ohnehin nur Zahlenspielereien. Großstadtkammern mit einem sozusagen anonymen Prüfungsklima haben schlechtere Ergebnisse als Kammern mit z. B. 300 km Längenausdehnung, wo sich das Prüfungsgeschehen in vielen kleinen Städten abwickelt und zudem die Auslese durch weiterführende Schulen weniger wirksam ist.

Die früher üblichen Bewertungsrichtlinien (Empfehlungen der Kammer an ihre Prüfungsausschüsse ohne rechtliche Verbindlichkeit) werden zunehmend durch neue Ausbildungsordnungen mit rechtsverbindlichen Bewertungsvorschriften abgelöst, die von Beruf zu Beruf neue Varianten einführen; in der Prüfungspraxis ließ das schon das böse Wort von einer Balkanisierung der Prüfungsvorschriften aufkommen.

Daß auch Schulnoten keinen exakten regionalen Vergleich zulassen, ergibt sich aus der Misere der Abiturientennoten für die Hochschulzulassung. Hier muß die Ständige Konferenz der Kultusminister unerklärliche Länderunterschiede durch einen Bonus oder Malus ausgleichen.

Jedes noch so logisch oder unlogisch konstruierte Bewertungssystem läuft auf eine subjektive fachmännische Schätzung hinaus. Es gibt weder für Schul- noch für Berufsprüfungen ein Bewertungssystem, das objektiv genau funktioniert und absolute Gerechtigkeit garantiert. Ein Test bei Industriefacharbeiterprüfungen (aber nicht nur hier!) würde das schnell beweisen: Man brauchte nur ein bestimmtes Prüfstück und die hierfür verwendete Zeit von mehreren Prüfern getrennt bewerten zu lassen: mit Sicherheit kämen differenzierte Bewertungen zustande. Der Garant für eine vernünftige Bewertung ist die Person des Prüfers, sein fachmännisches Urteil und sein gesunder Menschenverstand.

#### **Die hundert Punkte (50 — 17 — 14 — 11 — 8)**

Einiger gemeinsamer Nenner aller Bewertungsregulative für Abschlußprüfungen in rund 500 anerkannten Ausbildungsberufen der Industrie- und Handelskammern, Handwerkskammern und anderen „zuständigen Stellen“ ist die Anlehnung an das sechsstufige Notensystem der (Berufs-)Schulen und die Feinaufteilung in ein Punkte-Bewertungssystem. Die Verwendung der Schulnoten ist nicht umstritten, anders dagegen die 100 Punkte, denen dieser Aufsatz gewidmet ist und nachweisen will, daß

- sechs Noten, gegebenenfalls mit Zwischennoten, für nahezu alle Bewertungsakte ausreichen;
- gegebenenfalls den 6 Noten Dezimalstellen angefügt werden können, womit ein System von 60 Punkten entsteht, das sich den sechs Schulnoten logisch und mit gleichen Abständen anpaßt; es wäre rechnerisch einfach zu handhaben und leidlich praktikabel. Dieses System ist für Abschlußprüfungen nirgendwo existent und sogar seit 1971 rechtlich nicht statthaft;
- das gegenwärtig geltende 100-Punkt-System prinzipiell und erst recht mit seinen ungleichen Punktabständen unlogisch ist;
- Multipunktsystem nur scheinbare Genauigkeit bieten
- und Multipunktsysteme daher nicht gerechter und damit vergleichbarer bewerten lassen. Im Gegenteil: sie sind stärker manipulierbar.

Der Fachmann bewertet eine praktische Prüfungsleistung zunächst einmal als „brauchbar“ (dem Abnehmer des Produktes zumutbar) oder „die Arbeit kann noch nachgebessert werden“ oder als „Ausschuß“. Das kann er noch etwas differenzieren. Allmählich wurden die Bewertungsvorschriften für Berufe mit praktischen Prüfstücken und Arbeitsproben immer komplizierter und die Modelle zahlreicher. Auch hier wurde deutlich, daß der Teufel im Detail steckt.

Nach langen Bemühungen um ein einheitliches System verständigten sich die Industrie- und Handelskammern im Kompromißwege auf das zu diesem Zeitpunkt bereits überwiegend praktizierte System 50 — 17 — 14 — 11 — 8. Die Musterprüfungsordnung des Bundesausschusses für Berufsbildung übernahm es 1971. Die Kammern beschlossen diese Musterprüfungsordnung als Satzungsrecht, so daß dieses Punktschema eine größere Verbindlichkeit hat als die bisherigen Bewertungsrichtlinien. Die Musterprüfungsordnung sieht in § 20 (Bewertung) folgendes Schema vor:

Noten	Punkte	Punktabstände	Zum Vergleich 60-Punkt-System
1	100—92	8	bis 60
2	unter 92—81	11	bis 50
3	unter 81—67	14	bis 40
4	unter 67—50	17	bis 30
nicht bestanden			
5	unter 50—30	20	bis 20
6	unter 30	30	bis 10

Jeder vernünftige Prüfer wird beim einzelnen Bewertungsakt nur volle Punkte geben, ohne daß § 20 dies ausdrücklich vorschreibt. Andererseits entstehen unvermeidlicherweise Punkte mit Dezimalstellen durch Faktorenrechnung oder durch Sammelbewertung von mehreren einzelnen Punktpositionen. Damit entsteht das Problem der Auf- und Abrundung, zu dem § 20 keine klare Auskunft gibt. Die Kategorie „unter 92 bis 81“ zum Beispiel scheint anzudeuten, daß eine knapp unter 92 liegende Punktzahl (z. B. 91,8) nicht auf 92 und damit auf „sehr gut“ aufgerundet werden darf.

Die höchst sonderbare Entstehungsgeschichte soll nach Mitteilung des inzwischen verstorbenen Dipl.-Ing. Hefer, seinerzeit IHK Hagen, folgende sein: Ausgangspunkt der Autorengruppe war die Gauß'sche Häufigkeitskurve. Sie fußt auf der Beobachtung, daß bei einer ausreichend großen und ausreichend gemischten Gruppe von Menschen — vereinfacht ausgedrückt — eine Leistungsmessung häufig ein Minimum sehr schlechte und sehr gute, ein Maximum unter-, mittel- und überdurchschnittliche Leistungen erbringt. Dies läßt sich zeichnerisch in Form einer Glockenkurve darstellen. Man teilt offenbar diese Glocke in sechs senkrechte Segmente, die man sich als eine Entsprechung der sechs Schulnoten dachte und zwar so, daß auf Segment 1 8 % aller Gruppenangehörigen entfallen, auf Segment 6 ebenfalls 8 %. Eine solche Einteilung läßt sich (laut Hefer) berechnen mit der Formel

$$\text{Punktzahl } y = 100 - 1,5 x^2 - 6,5 x$$

wobei x die Note ist. Setzt man die Note x = 2, ergibt sich die Punktzahl y = 81 Punkte als Untergrenze die Note 2; setzt man die Note x = 4, so ergibt sich die Punktzahl y = 50 Punkte als Untergrenze der Note 4.

Vielfach ist man der irrtümlichen Meinung, daß dieses System, das für „sehr gut“ nur 8 Punkte vorsieht, damit auch die Zahl der Absolventen mit „sehr gut“ in angemessenem Rahmen hält. Das ist u. a. schon deshalb falsch, weil auch ein 6- oder 60-Punkt-System bei leidlich korrekter Anwendung die gleiche Streuung bietet, es sei denn, daß spezielle Bewertungsvorgaben insbesondere bei Prüfstücken (z. B. Funktion, Passung, Oberflächengüte) den Prüfungsausschuß geradezu zwingen, massenhaft gute Teilnoten zu vergeben.

Der andere, sozusagen nachträglich eingebrachte Grundgedanke dieses Punktsystems 50 — 17 — 14 — 11 — 8 war, wie dem Verfasser aus zahlreichen Beratungen seit 1952 erinnerlich ist, daß die erreichte Punktzahl den Prozentsatz der erwarteten Voll-Leistung bestätigen soll. Das könnte durchaus einmal zufällig richtig sein, aber grundsätzlich ist es wohl ein Irrtum, anzunehmen, daß z. B. 55 Punkte gleichzusetzen wären mit 55 % der von einem Gehilfen erwarteten Voll-Leistung. Weder steht diese Voll-Leistung immer eindeutig fest, noch läßt sich die Leistung des Prüflings — vielleicht von Ausnahmen abgesehen — eindeutig zu dieser (fiktiven oder feststellbaren) Voll-Leistung in Beziehung setzen.

Ein Hundert-Punkt-System ist keineswegs leistungsgerechter als ein einfaches 6-Noten-System, denn:

1. Prüfungsleistungen, so z. B. mündliche und schriftliche Prüfungen, sind nur sehr begrenzt exakt meßbar. Dem-

entsprechend kamen die Kammern — jedenfalls bis 1971 — in den Kaufmannsgehilfenprüfungen ganz gut ohne Punktbewertungsrichtlinien aus. Offenbar ist es also die Mentalität des Prüferstabes in den Facharbeiterprüfungen, die hinter dem Willen zur genauen Punktbewertung steht. Dieser Personenkreis ist gewohnt, mit Tabellen, Zeichnungen und Feinmaßstäben umzugehen.

2. Die Prüfer gehen daher meist von einer Note aus und sehen dann in der Tabelle nach, welcher Punktraum dieser Note zugeordnet ist. Wenn zum Beispiel in einem Druckberuf eine Farbmischung zu bewerten ist, versagen Punktmaßstäbe. Mit anderen Worten, eigentlich funktioniert solcherart auch das dummste Punktsystem.
3. Die Manipulierbarkeit eines Hundert-Punkt-Systems ist größer als diejenige einfacher Noten. Viele Prüfer benutzen den einer Note vorgegebenen Punktraum nicht im Sinne einer exakteren Bewertung, sondern um dem Prüfling in einer Grenzlage zu helfen. Beispiel: Ein Prüfling erreicht in einem wichtigen Prüfungsteil nur 49 Punkte. Sonderbarweise kommt er gar nicht so selten bei einer weiteren für das Bestehen wichtigen Prüfungsaufgabe auf 52 Punkte; oder der Prüfer korrigiert eine bereits gegebene Punktzahl nachträglich.

Es kommt hinzu, daß erfahrungsgemäß Beschreibungen von Prüfungsstücken (gleichgültig, ob bezirks-, landes- oder bundeseinheitlich erstellt) trotz besten Bemühens der beteiligten Fachleute nicht immer ganz frei von Mängeln (z. B. falsche Beschriftungen, überholte DIN- oder Bearbeitungszeichen, Fehler in der Zeichnung, in der Werkstoff- und in der Werkzeugliste, fehlerhafte Schaltzeichnungen, nicht angemessene Zeitvorgaben etc.) sind.

#### Sonstige Fragen

Auch ein 100-Punkt-System enthält gewöhnlich berufsbezogen noch spezielle Elemente, z. B. werden einzelne Prüfungsfächer mit Multiplikatoren versehen, um sie als wichtiger herauszustellen. Auch spielt der Gedanke eine Rolle, man müsse die Arbeitzeit des Prüflings bewerten, gegebenenfalls bei Arbeitsproben (im Gegensatz zu Prüfungsstücken, die nur abschließend bewertet werden) das Arbeitsver-

halten. Aus langer, leidvoller Erfahrung weiß der Verfasser darüber zu berichten, daß diese zusätzlichen Einzelheiten immer unterschiedlicher wurden und keinen Nutzen erbringen.

1. **Die Gewichtung einzelner Prüfungsaufgaben** mit Multiplikatoren (Faktoren) hat sich bei einzelnen Durchrechnungen als überflüssig erwiesen, weil sie die Prüfungsresultate nur ganz geringfügig verändert. Außerdem bietet sie Anlaß zu Rechenfehlern. Bei jeder neuen Ausbildungsordnung ist das Experimentieren mit diesen Multiplikatoren eine Versuchung für die betreffenden Fachleute, immer wieder neue Varianten zu erfinden.
2. **Spezialtabellen:** Wo genau meßbare Leistungen vorliegen, können Spezialtabellen dienlich sein. Hier findet sich nirgends das 100-Punkt-System, sondern es werden eigenständige Maßstäbe verwendet und 6 Noten zugeordnet, z. B. glatter Satz oder Diktat beim Schriftsetzer nach Fehlern, Kurzschrift und Maschinenschreiben nach Silben und Anschlägen.
3. **Arbeitszeit:** Jede Prüfungsaufgabe enthält eine Sollzeitvorgabe. Der Prüfling kann gegebenenfalls langer oder kurzer arbeiten. Hier sollte man auf lange Spezialtabellen zur Arbeitszeitbewertung verzichten. Erhebliche Zeitüberschreitungen eines Prüflings innerhalb einer größeren Gruppe sollte der Prüfungsausschuß nach freiem Ermessen bewerten. Dies auch deshalb, weil die beim konstruktiven Entwurf der Prüfungsstücke vorgegebene Auffertigungszeit (Sollzeit) ständig umstritten ist, selbst wenn sie in Betrieben getestet wurde.
4. **Der Bundesverordnungsgeber** sollte es in Zukunft unterlassen, in jeder neuen Ausbildungsordnung neue Varianten der Bewertung einzuführen, es sei denn, daß sich solche aus der Art des Berufes als notwendig erweisen. Insbesondere sollte die zusätzliche mündliche Prüfung nach Wunsch des Prüflings, die nur verbessern kann, entfallen. Sie ist eine gut gemeinte Wohlfahrtslösung, die mathematisch exakt überhaupt nicht zu bewerten ist. Sie erinnert an einen Länder-Zehnkampf, in dem man ver einbaren möchte, daß die unterliegende Mannschaft einen 11. Wettbewerb nachschieben darf.

## Noch einmal: 2jährige Metallberufe

#### Stellungnahme des Gesamtverbandes der metallindustriellen Arbeitgeberverbände e. V.:

In einem Aufsatz „Differenzierung des Ausbildungsangebots-Instruments zur Minderung des Ausbildungsplatzmangels?“ befaßt sich Dagmar Lennartz in BWP 1/77 kritisch mit dem Vorschlag des Gesamtverbandes der metallindustriellen Arbeitgeberverbände (Gesamtmetall), bei der anstehenden Neuordnung der Metallberufe neben Ausbildungsberufen mit 3jähriger Ausbildungsdauer zusätzlich fertigkeitsbetonte 2jährige Ausbildungsgänge einzuführen, die nach Auffassung des Verbandes weniger begabten Jugendlichen eine Ausbildungschance in den Metallberufen eröffnen sollen. Gesamtmetall nimmt zu der Analyse und Bewertung wie folgt Stellung:

1. Die Arbeitgeber der Metallindustrie bekennen sich bewußt zu einer beschäftigungsorientierten Berufsausbildung. Berufsausbildung manifestiert und legitimiert sich in erster Linie dadurch, daß sie den jungen Menschen eine seinen Begabungen und Fähigkeiten angemessene Berufs- und Einkommenschance eröffnet. Dies gilt insbesondere in Zeiten steigender Jugendarbeitslosigkeit. Die zielgerichtete Beschaf-

tigungsorientierung ist bei den vorgeschlagenen 2jährigen Ausbildungsgängen um so vertretbarer, als im 1. Ausbildungsjahr die gleiche breite berufliche Grundbildung vermittelt werden soll, wie für die 3jährigen Ausbildungsberufe. Die „Mobilitäts-Ausbildung“ wäre also in beiden Ausbildungsgängen gleich

2. Die Auffassung, daß bei einer „differenzierten Ausbildung“ das Prinzip der Durchlässigkeit faktisch aufgehoben sei, ist falsch. Wir haben nie einen Zweifel daran gelassen, daß den Absolventen 2jähriger Ausbildungsberufe der Übergang zu den 3jährigen Ausbildungsberufen möglich sein müsse, wenn die entsprechenden Voraussetzungen vorliegen. Das Vorliegen dieser Voraussetzungen müßte anhand von sachgerechten Kriterien überprüfbar sein. Demgegenüber sind wir der Auffassung, daß in der Stufenausbildung die Durchlässigkeit de facto aufgehoben ist. Da der Übergang von der 1. zur 2. Stufe von keinerlei Eignungskriterien abhängig gemacht wird, andererseits aber die Ausbildungsbetriebe die volle Verantwortung für den Erfolg der Ausbildung tragen, ist z. B. bei der Stufenausbildung in den Elektroberufen zu beobachten, daß die Ausbildungsbetriebe nur noch Jugendliche