

## Die Bewertung von schriftlichen Prüfungen im Fach Deutsch („Aufsätzen“)

### 1. Rechtliche Ausgangslage:

Es gibt – außer den Vorschriften zu Gewichtung von schriftlichen Leistungsnachweisen - keine gesetzlichen Vorgaben zur Bewertung von Aufsätzen (außer RSO §41; siehe → Erläuterungen zum Korrekturblatt).

### 2. Hauptgütekriterien von Tests und ihrer Beurteilung:

- **Objektivität:** Bewertung unabhängig vom Korrektor und den Bedingungen (Tageszeit etc.) → Ideal:  
Verschiedene Korrektoren kommen zu verschiedenen Zeiten (=Durchführungsobjektivität/Bedingungskonstanz) zum gleichen Ergebnis (=Auswertungsobjektivität), das sie auch gleich beurteilen (Interpretationsobjektivität; eine 1 ist nicht bei einem perfekt und beim anderen nahezu perfekt)  
Bei der Beurteilung wichtig: *Auswertungs- und Interpretationsobjektivität*
- **Reliabilität:** „Unter der Reliabilität eines Tests versteht man den Grad der Genauigkeit, mit dem er ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal misst.“ (Lienert, 1969). Meist wird darunter *Wiederholungsreliabilität* verstanden = Wiederholung der Prüfung führt unter gleichen Bedingungen zu gleichen Ergebnissen
- **Validität:** „Die Validität eines Tests gibt den Grad der Genauigkeit an, mit dem dieser Test dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die) er messen soll oder zu messen vorgibt, auch tatsächlich misst. Ein Test ist demnach vollkommen valide, wenn seine Ergebnisse einen unmittelbaren und fehlerfreien Rückschluss auf den Ausprägungsgrad des zu erfassenden Persönlichkeits- oder Verhaltensmerkmals zulassen, wenn also der individuelle Testpunktwert eines Probanden diesen auf der Merkmalskala eindeutig lokalisiert.“ (Lienert, 1969)

Meist wird darunter die sog. die Kriteriums-/Konstrukt- oder Inhaltsvalidität gemeint:

- *Inhaltsvalide* Tests spiegeln das zu Messende unmittelbar wieder (Tastschreiben nach Geschwindigkeit wird getestet durch einen Schnellschreibetest, Kopfrechenleistung durch Aufgaben, die nicht aufgezeichnet sind)
- *Kriteriums- und konstruktvalide* Tests beziehen sich auf andere Tests oder theoretische Konstrukte. Messen Schulnoten die Intelligenz? Oder: Misst ein Deutschaufsatz das im Lehrplan angegebenen Bildungsziel bzw. die im Bildungsstandard formulierten Kompetenzlevel (einen Text verstehen durch Analyse und Stellungnahme in schriftlicher Form). Misst der Test nicht auch irrelevante Kriterien (Affektkontrolle, Angstbewältigung, Stresskontrolle, motorische Fähigkeiten, Konzentrationsleistung etc.)  
Meist: Tatsächliche Messung des curricular Vorgeschriebenen und im Unterricht Erarbeiteten

### 3. Auswahl wichtiger Funktionen von Leistungsmessungen

- Informations- und Rückmeldefunktion: Auskunft über Leistungsstand (wo stehe ich/er/sie?)
- Kontrollfunktion durch Lernerfolgskontrolle, Erfolgskontrolle für SchülerInnen wie LehrerIn (was muss ich/er/sie wo noch tun?)
- Vergleichsfunktion für Schüler und Lehrer (Vorsicht!)
- Erziehungs- und Disziplinierungsfunktion (ist verboten!)
- Motivation: Motivation durch gute Leistungen, aber auch Motivation durch Leistungsschwächen
- Selektions- und Allokationsfunktion (Zuweisung zu Schultypen, Klassen nach Notenrang)
- Prognosefunktion (Berufswahl, Lebensplanung, Karriere?)

### 4. Qualitätskriterien gerechter Leistungsbewertung:

### Allgemeine Bemerkungen zu Ziffernnoten

Ziffernnoten ergeben ein *Ordinal- oder Rangskala*, die Leistungen nach dem Grad der Ausprägung eines Merkmals einstufen.

Das Merkmal kann dabei sehr verschieden sein (Aufsagen von Wörtern, Begriffen, Erklären von Begriffen, Lösen von Problemen, Formulieren eines Sachverhalts, Treffen eines Ausdrucks, Schreibung eines Wortes usw.).

Das Rangkriterium zwischen Noten ist nur  $\leq$  oder  $\geq$ , aber nicht selbst quantifiziert. Eine Leistung mit Note 1 ist nicht sechsmal besser als eine 6, es gilt nur *Transitivität bzgl. „besser“ oder „schlechter“*: Eine Leistung mit 1 muss besser sein als die einer 2, die wiederum besser sein muss als eine 3 usw. bis 6.

Es gilt *keine Reflexivität*: zwei 1er müssen nicht gleich gut sein!

→ Die Abstände zw. Noten sind relativ und können schwanken: Leistungsdifferenz zw. 1 und 2 kann kleiner sein als zw. 4 und 5 (ist es auch in der Regel!).

Ziffernnoten geben nicht notwendig eine *Intervall* (gleiche Intervalle wie bei Temperaturskala: 2 °C ist nicht doppelt so heiß wie 1°C, stellt aber das gleiche Intervall dar wie zw. 3 und 4°C)- oder *Proportionalskala* (gleiche absolute Abstände wie Waage: 2 kg ist doppelt so schwer wie 1 kg) wieder. Bsp. Aufsatz: Bereich Sprache 0-1 F = Note 1; 2-4 F = 2; 3-5 F = 3

→ Abstand zw. 1 und 2 geringer als zw. 2 und 3

In der Mathematik oft besser durch linearen Schnitt: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60P

### 4.1 Differenzierungsfähigkeit:

Messung unterschiedlicher Leistungsfähigkeit

### 4.2 Maßstäbe und Bezugsnormen bei der Bewertung:



#### 4.2.1 Frageschema

WEN

- Alter: Vorwissen
- Klassenstufe: Lehrplan

WAS

- Reproduktion
  - Transferleistungen
  - kreativer Umgang
- (jeweils harte [wie Umfang, Aufbau, Tempus] und weiche Maßstäbe)

WOZU

- Fähigkeiten
- Fertigkeiten

#### 4.2.2 Idealtypen von Bezugsnormen:

- Soziale Bezugsnorm (relative Bewertung): Bewertung eines Aufsatzes relativ zum Leistungsstand der Klasse.
- Kriteriumsbezogene Bezugsnorm (absolute Bewertung): Bewertung nach einem objektivierbaren oder sogar objektiven Schlüssel (Kriteriumskatalog).
- Individuelle Bezugsnorm (individuelle Bewertung): Bewertung am Potenzial des Schülers.

### 5. Der Bewertungsprozess:

Qualitative Bewertung („Finden“) → Quantifizierung („Zählen“) → Qualitative Bewertung („Gewichten“) → Quantifizierung („Note“)

Faustregel: *Je formaler das Einzelkriterium, desto besser lässt es sich quantifizieren: Rs/Zs → Sprache → Inhalt*



### 6. Das Bewertungsverfahren im Einzelnen:

#### 6.1 Teilbereiche

Für die Aufsatzbeurteilung hat sich die Bewertung nach drei Teilbereichen „bewährt“:

- a) Inhalt: Aufbau und Durchführung
- b) Sprache: Wortwahl, Satzbau (weitere Unterbereiche je nach Aufsatzart)
- c) Formales: Rechtschreibung, Zeichensetzung



#### 6.2 Gewichtung.

Abhängig von...

- von Schwerpunkten im Unterricht: Was wurde zuvor mehr geübt?
- der Aufsatzart, z.B. *betont eine Schilderung mehr den sprachlichen Ausdruck*.
- dem Schwierigkeitsgrad des Aufsatzes: *Je mehr konzeptionelle Fähigkeiten gefragt sind, desto mehr sollte der inhaltliche Teilbereich gewichtet werden.*



- der Klassenstufe: *Je höher die Klassenstufe, desto weniger sollte das Formale gewichtet werden.*

### 6.3 Probleme der Bewertung der verschiedenen Teilbereiche:

#### a) Der Wiederholungsfehler in der Rechtschreibung:

- Definition 1: Fehler einer Gruppe, die erkennen lassen, dass eine Regel nicht richtig verstanden wurde bzw. nicht richtig angewandt wurde. Klassiker ist das falsche Dehnungs-h: sie wa~~r~~en, sie ka~~m~~en.

- Definition 2: Fehler bei denselben Wörtern.

b) Analoges gilt bei der Bewertung des sprachlichen Ausdrucks (z.B. wiederholte falsche Satzgliedstellung bei den selben Satzgliedern) und des Inhalts (z.B. Verwechslung von Begründung und Beispiel in der Erörterung; dauernder Bezug der Rückführung auf die Themenfrage statt auf die These)



Grobe Faustregel: *Ein Wiederholungsfehler wird aufgehoben, wenn die fragliche Regel/Norm mindestens einmal richtig angewendet/beachtet wurde.*

### 6.4 Auswahl von „Traps“/Beurteilungsfehlern



#### 1.) Effekte, in denen Eigenschaften übertragen werden

- Rosenthal-Effekte: Verhalten → Leistung. Vorurteile gegenüber Schülern (Studie von Weiss, 1995): Miteinfließen von Bewertungen durch Kenntnis des Schülers (Mitarbeit wie Verhalten im Unterricht, „Lieblingsschüler“, Umfeld des Schülers) → Name abdecken
- „Halo-Effekte“: Merkmal 1 → Merkmal 2. Eine Eigenschaft wird auf einen anderen Bereich übertragen: ein rechtschriftlich guter Aufsatz/Schüler wird auch als sprachlich gut eingeschätzt
- Stereotypisierungen: Generelle Leistung → spezielle Leistung. Schlechter Schüler → schlechter Aufsatz, weil nur negative Eigenschaften wahrgenommen werden.
- Logische Fehler: Bereichsleistung → spezielle Leistung. „Ein in einem sprachlichen Fach guter Schüler muss auch in Deutsch gut sein“ (kausale Fehlattributierung aufgrund einer invarianten, stabilen, internalen Sprach-Begabung)

#### 2.) Zeitbezogene Effekte

- Proximity-Errors: Unmittelbar hintereinander beurteilte Merkmale oder Teile wechselwirken. Teil A gut → Wahrscheinlichkeit, dass Teil B als gut bewertet wird, steigt.
- Primacy-Recency-Effekt: Als besonders wichtig wird der Anfang und das Ende einer Leistung beurteilt. Ein schlechter Beginn führt zu Proximity-Errors (o weia, das geht ja schon „gut“ los!) und umgekehrt kann eine gute Leistung durch einen katastrophalen Schluss konterkariert werden.

#### 3.) Fehler-Effekte

- Quantifizierungs- oder Schematisierungsfalle: Über den Aufsatz wird eine Art detailliertes Lösungsraster (mit Fehler- oder Notenschlüssel) gelegt und dann jeder Punkt nach einem binären Schema bewertet („hat erfüllt“, „hat nicht erfüllt“). Problem: Es gibt nur Eins und Sechs für jeden Punkt. Versagt vollkommen bei kreativen Aufsatzformen und oft bei der Bewertung inhaltlicher Kriterien.
- Fehler-Fehler-Falle:
  - Bewertet wird nach dem Maßstab, dass die Bestnote irgendetwas mit Fehlerfreiheit zu tun hat (siehe → Erläuterungen zum Korrekturblatt).
  - Unberücksichtigt bleiben die positiven, besonders gute Punkte. (Randbemerkungen „Gut!“, „Sehr gut!“ usw.)

#### 4.) Beurteilungstendenzen, die von der Persönlichkeit, aber auch externen Faktoren abhängen

- Tendenz zur Milde. Gründe: Sympathie bei Schülern, Angst vor Einwänden und Kontrolle
- Tendenz zu Extremen. Gründe: (falsch verstandene) Motivation, unvollständiges Raster
- Tendenz zur Strenge: Autorität zeigen, zu hohe Anforderungen (Kriteriumskatalog!)

### 6.5 Die Bemerkungen:

#### 6.5.1 Randbemerkungen:

Sie helfen SchülerInnen oft mehr als die Verbalbeurteilung, da sie hier einen konkreten Bezug zum Selbstverfassten sehen. Standardbemerkungen: Unklar!, Ungenau!, Sinn?, Satz!

#### 6.5.2 Die Verbalbeurteilung und ihre Kriterien:



- muss individuell und darf nicht allgemein sein (nicht bei allen dasselbe Schema: „Dein Inhalt ist mangelhaft, deine Sprache ausreichend, deine Rechtschreibung gut“)
- muss/sollte Bezug nehmen auf die im Unterricht erarbeiteten Aufsatzkriterien
- hat hilfreich zu sein und soll konstruktive Kritik bieten. Hilfreich bedeutet: Dem Schüler auch konkrete Handlungsanweisungen und Tipps geben („Achte noch mehr auf xy und wiederhole die Regeln xy“; „Bemühe dich um...“). Konstruktiv heißt: Den Schüler nicht „in die Pfanne hauen“, auch wenn Fehler sich in jedem Aufsatz wiederholen (Motivationskriterium beachten!)
- sollte die Gesamtnote für den Schüler nachvollziehbar machen (Gewichtung von negativer und positiver Kritik auf Note beziehen)
- muss konkrete Bezüge zum Aufsatz enthalten (Zitate!)
- darf und sollte auch Bezug nehmen auf die individuelle Entwicklung („Dein Satzbau hat sich enorm verbessert“, „Deine Rs hat sich sehr verschlechtert“)
- kann auch persönliche Einschätzungen enthalten („leider“, Grenzkommentar: „ärgerlich“), sollte sich jedoch pädagogischer Bemerkungen enthalten (Ausnahmen zugelassen).

#### Literatur:

Amelang, Manfred / Bartussek, Dieter / Stemmler, Gerhard / Hagemann, Dirk (Hgg.): Differentielle Psychologie und Persönlichkeitsforschung, Stuttgart: Kohlhammer, 6. Aufl. 2006.

Formen der Leistungserhebung im Fach Deutsch. Eine Handreichung für die bayerischen Realschulen zu den Möglichkeiten der Leistungserhebung und -bewertung im Fach Deutsch unter besonderer Berücksichtigung neuerer Unterrichtsformen, hg. v. ISB Bayern, Donauwörth 2005.

Ingenkamp, Karlheinz (Hg.): Zur Fragwürdigkeit der Zensurgebung. Weinheim: Beltz, 9. Aufl. 1995.

Lienert, Gustav A.: Testaufbau und Testanalyse, Weinheim: Beltz 1969.

Lukesch, Helmut: Einführung in die pädagogisch-psychologische Diagnostik, Regensburg: Roderer 2002.

Rheinberg, Falko: Bezugsnormen und schulische Leistungsbeurteilung, in: Franz E. Weinert (Hg.): Leistungsmessungen in Schulen, Weinheim: Beltz 2002.

Ulrich, Klaus: Einführung in die Sozialpsychologie der Schule, Weinheim: Beltz 2001.

Weiterhin relevant: Studien von Hartog & Rhodes 1936, Dicker 1977 und Ascherleben 1971