Testtheoretische Kritik mündlicher und schriftliche Prüfungen (29.5)

- Messung und Notengebung
- Kritik an schriftliche Prüfungen
- Kritik an mündlichen Prüfungen
  - Optimierung mündlicher Prüfungen

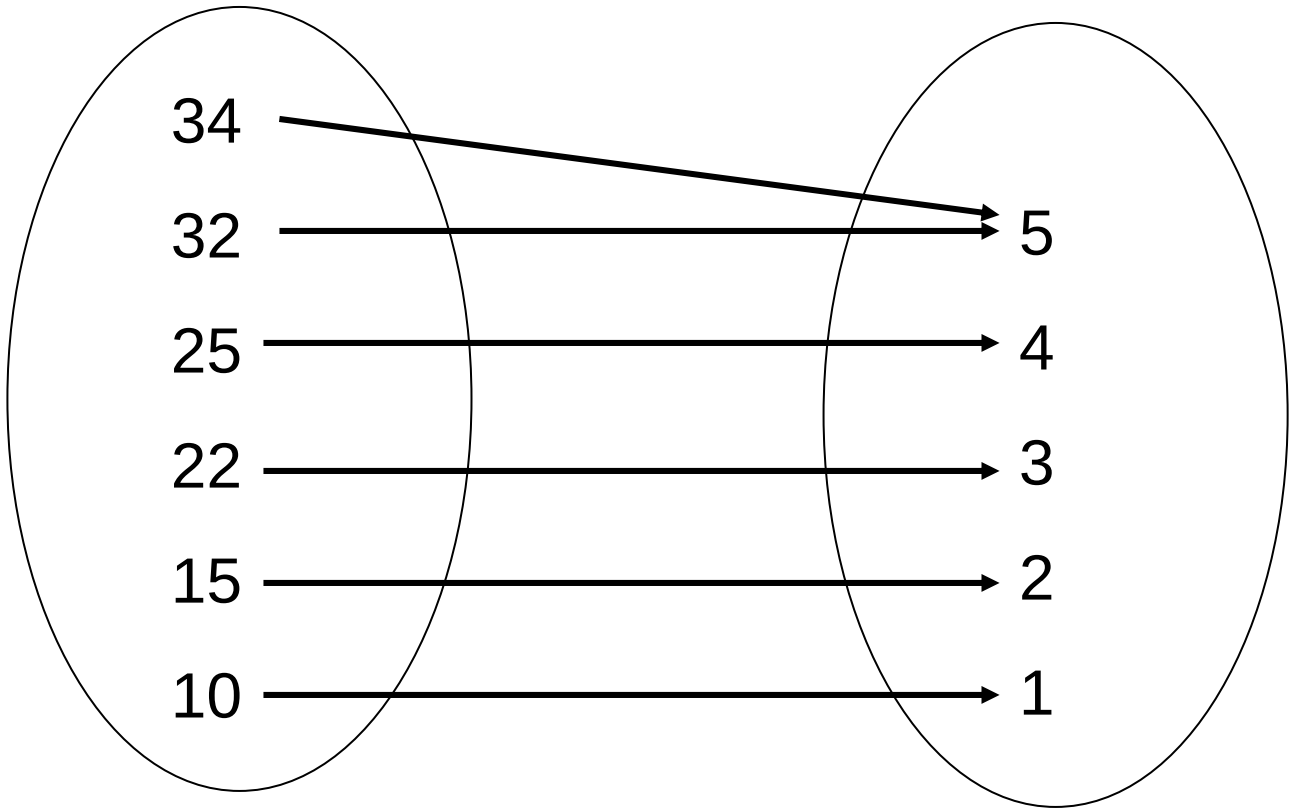
Messung

Objekte

z.B. Diktatfehler von 6 Schülern

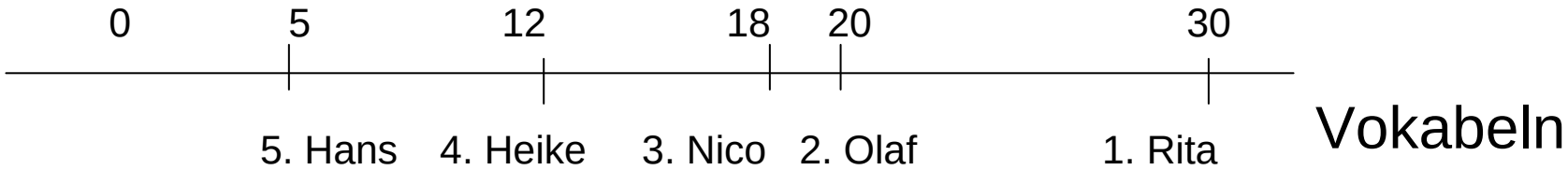
Zahlen

Rangplätze nach Leistungsgüte



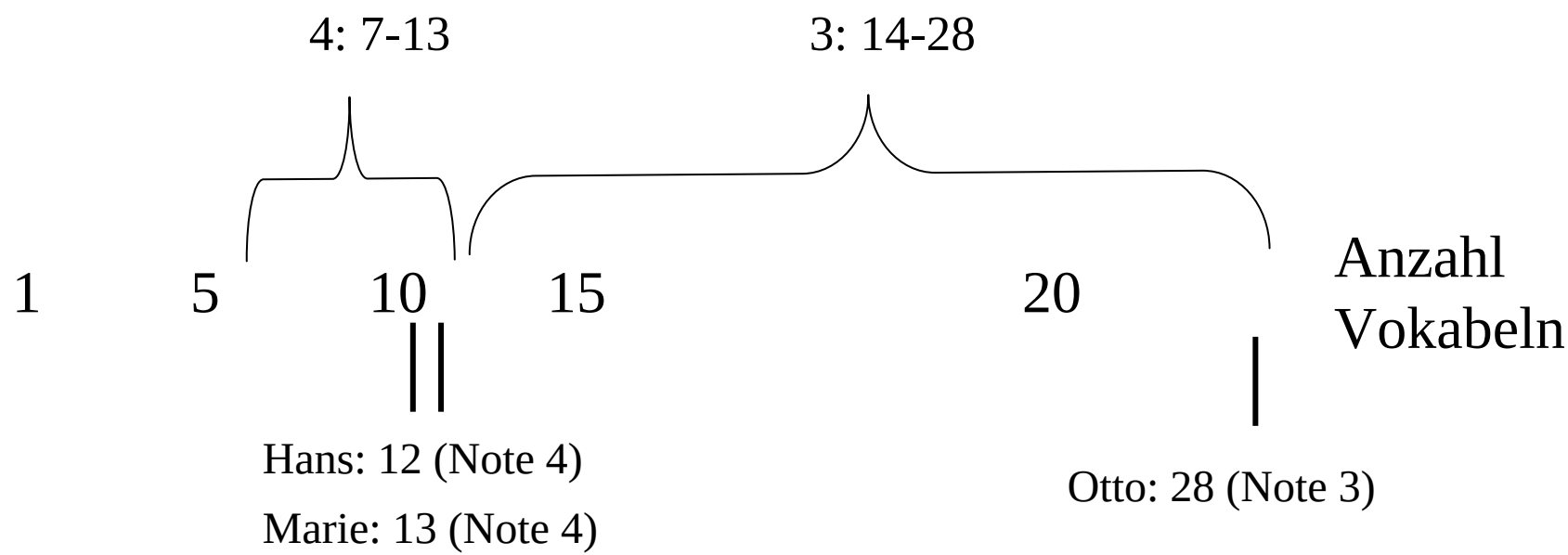
Die 4 Niveaus der Messung

- Nominalskala: Gleichheit vs. Verschiedenheit der Zahlen repräsentieren Gleichheit vs. Verschiedenheit der Objekte hinsichtlich eines Merkmals (z.B. Geschlecht, Schulklasse)
- Ordinalskala (= Rangskala): Reihenfolge (Rangfolge) der Zahlen (z.B. **Noten**) repräsentieren Rangfolge der „Objekte“ (-> keine Mittelwerte, Median sinnvoll)



- Intervallskala: Gleiche Abstände zwischen Zahlen repräsentieren gleiche Abstände der „Objekte“ (z.B. Celsius-Skala, aber: kein absoluter Nullpunkt, -> Mittelwert sinnvoll)
- Verhältnisskala: Proportionen (z.B. Gramm, Meter)

Ist Mittelwertsbildung innerhalb der Klasse sinnvoll?



Mittelwert Noten (11/3)= 3,7,->4 Mittelwert Vokabeln: (53/3) = 17,6-> Note 3

Fazit: Noten als Messung

- Lehrer ordnen Leistungsergebnissen (z.B. Fehleranzahl, Punktwerten usw.) Noten zu
- Diese Noten haben nur Rangskalen-Niveau, d.h. sie geben Informationen über die Rangreihe (der Fehleranzahl, der Punktewerte) innerhalb der Klasse
  - wegen des Klasseninternen Bezugssystems
  - Weil die Abstände zwischen Notenziffern nicht unbedingt den Abständen zwischen den „dahinter“ stehenden Leistungen entsprechen
- Daher ist streng genommen die Bildung von Mittelwerten von Noten innerhalb eines Schülers und zwischen Schülern nicht zulässig (besser: Median)

Der Median

- Noten: 1 1 1,5 2 6
- Mittelwert: 2,3
- Median: 1,5

Bezugsnormen und Funktionen der Leistungsbewertung

- Beurteilungsmaßstäbe: individuelle, soziale, kriteriumsorientierte Bezugsnorm (BnO)
- Funktionen von Beurteilungen (nach Ingenkamp, zitiert nach Lukesch, S. 447f.)
  - Schüler: Vergleich, Analyse/Selbstkontrolle, Anreiz
  - Lehrer: Analyse, Prognose, evtl. Selektion, Disziplinierung
  - Eltern: Vergleich, Bericht, Analyse, Prognose
- Bedeutung der Bezugsnormen für die Funktionen der Leistungsbewertung
  - Ind. BnO: v.a. Förderung, Motivierung (für Schüler)
  - Soz. BnO: v.a. Selektion, Berechtigung (für Lehrer, Gesellschaft)
  - Krit. BnO: v.a. Analyse, Bericht (für alle)

➢ Alle BnOs ergänzen sich, keine Einseitigkeit, Unterscheidung zwischen formeller Beurteilung und informeller Rückmeldung

Schriftliche Prüfungen

- Einstieg
- Objektivität
  - Auswertungsobjektivität
  - Interpretationsobjektivität
- Reliabilität
- Validität bzw. „sachfremde“ Einflüsse
- Verbesserungsmöglichkeiten

Objektivität schriftlicher Prüfungen (1)

- Objektivität (Auswertungsobjektivität: gleiche Arbeit, verschiedene Prüfer)
  - Starch & Elliot (1913): Examensarbeiten in Englisch, Mathematik, Geschichte und Mathematik werden von Lehrern (n= 180) deutlich unterschiedlich bewertet;
  - Hartog & Rhodes, (1936): Englischarbeit (von 48 Schülern) wird von ausgewählten Gutachtern deutlich unterschiedlich bewertet (trotz gleichem Auswertungsschema!)

Die Untersuchung von Weiß (1965):

Deutschaufsatz mit Vorinformationen

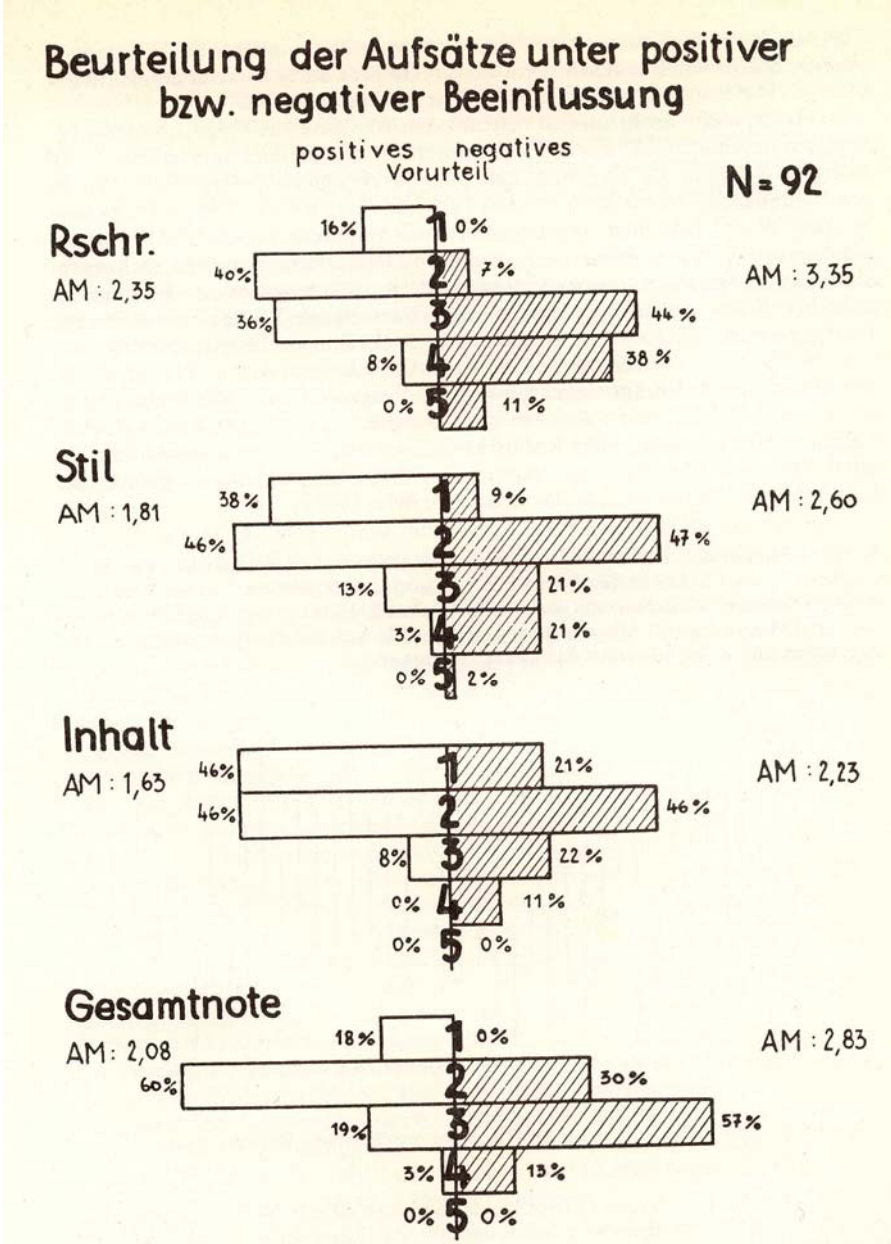


Abb. 12: Die Beurteilung von zwei Aufsätzen durch Lehrer (N = 92) unter positiver bzw. negativer Beeinflussung im Hinblick auf Rechtschreibung, Stil, Inhalt und Gesamtnote (AM = arithmetischer Mittelwert). (INGENKAMP 1977a, 113)

Objektivität schriftlicher Prüfungen (2)

- Objektivität (Auswertungsobjektivität: gleiche Arbeit, verschiedene Prüfer)
  - Lehrer unterscheiden sich nicht nur in den absoluten Noten, sondern auch in der Varianz vergebener Noten
  - Lehrer unterscheiden sich in der Differenzierung der Notengebung innerhalb eines Schülers über unterschiedliche Fächer
  - Ingenkamp (Tempelhofstudie): klasseninternes Bezugssystem

Schriftliche Prüfungen: das klasseninterne Bezugssystem

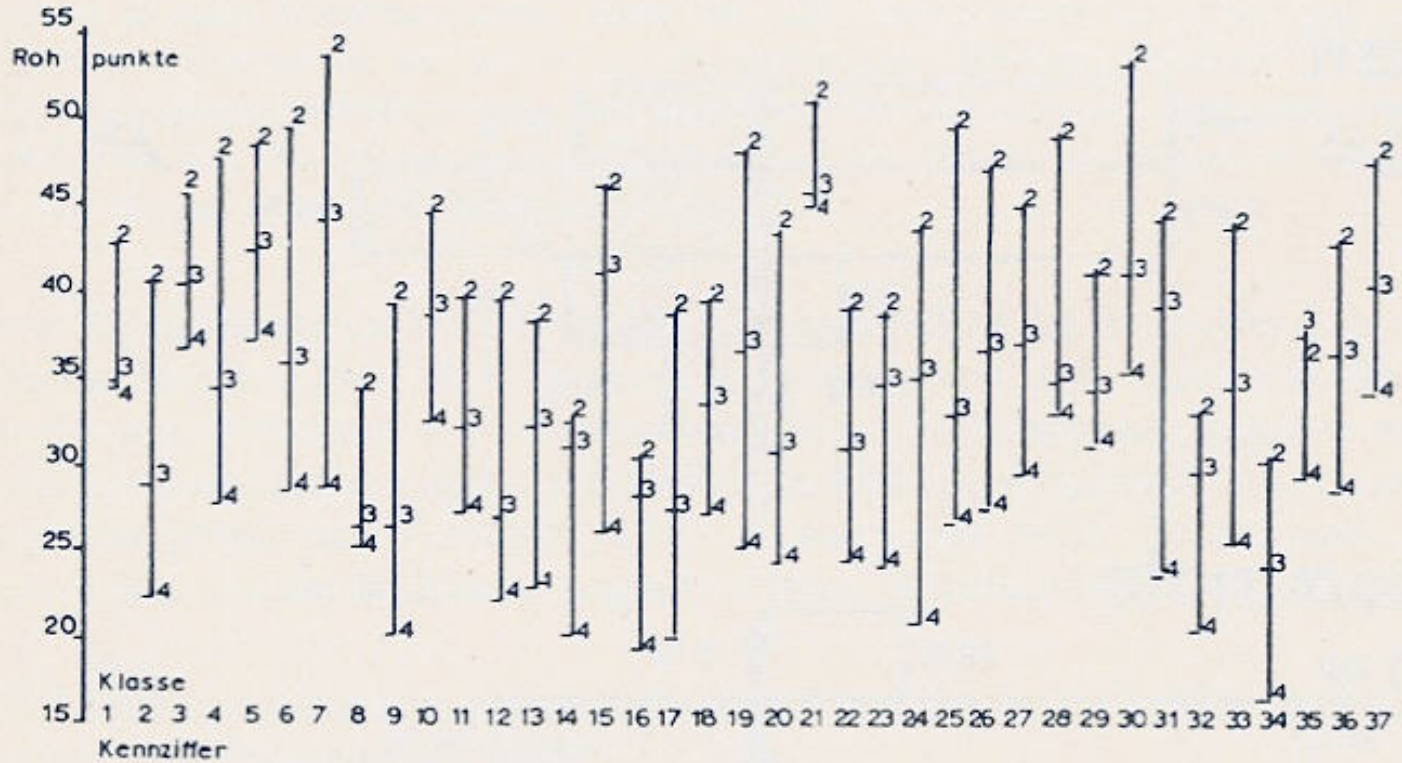


Abb. 13: Mittelwerte der Rohpunkte im HKI 8 bei verschiedenen Rechenzensuren (INGENKAMP 1977a, 197).

Reliabilität schriftlicher Prüfungen (2)

- Reliabilität (meist Wiederholungsreliabilität, gleiche Arbeit, gleiche Prüfer)
  - klassische Studie von Eels (1930): 61 Lehrer bewerten drei Kurzaufsätze aus Geographie und zwei Kurzaufsätze aus Geschichte innerhalb von 11 Wochen 2 mal (ohne Rückgriff auf Aufzeichnungen)

Die klassische Studie von Eels (1930)

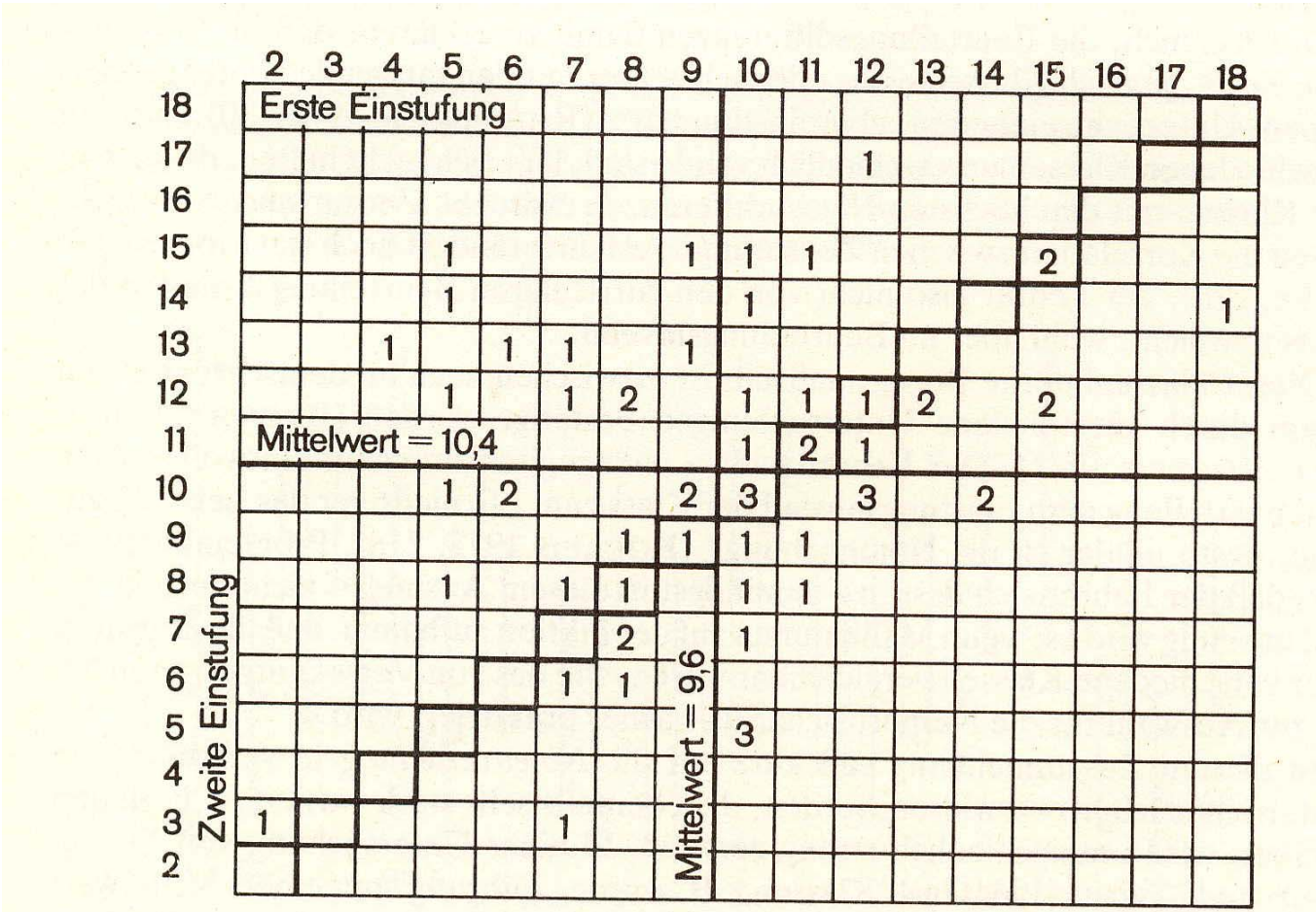


Abb. 14: Das Streuungsdiagramm zeigt die ersten und die zweiten Einstufungen ein Geographiearbeit durch dieselben Lehrer (INGENKAMP 1977a, 170)

Reliabilität schriftlicher Prüfungen (3)

- Reliabilität (meist Wiederholungsreliabilität, gleiche Arbeit, gleiche Prüfer)
  - Weitere Studien: Hartog & Rhodes, 1936 (nur Globalurteil „bestanden vs. nicht bestanden“; Finnlayson, 1951; Aschersleben, 1971, Dicker, 1973: Egal ob Mathematik oder Deutschsatz: Zeitstabilität und Paralleltestreliabilität (mehrere Arbeiten im gleichen Fach von einem Schüler) < .6 (aber große Schwankungen zwischen Lehrern)

Validität schriftlicher Prüfungen (1)

- Validität
  - Konstruktvalidität (sachfremde Einflüsse): Vorinformationen, soziale Stereotype (Weiss, 1965), Sozialschicht, Geschlecht (Hadley, 1954, Carter, 1952), Rechtschreibfehler und Sauberkeit, Fächer (in musischen Fächern bessere Noten), Bundesland, klasseninternes Bezugssystem, Klassengröße

Validität schriftlicher Prüfungen (2)

- Validität
  - Prognostische Validität: Grundschulnoten-Gymnasium: von ca. 16 (Undeutsch, 1960) bis .42 (Schenk-Danziger, 1963) und .45 (Roeder, 1997); Abiturnoten-Studienerfolg: ca .46 (Mathematik-Note bester Prädiktor)

Verbesserungsmöglichkeit bei schriftlichen Prüfungen

- Kriterienkatalog
- viele Einzelprüfungen (Reliabilität!)
- mehrer unabhängige Bewerter
- Speziell für Aufsätze: textganzheitliche und textanalytische Verfahren (Beck, 1979)
- Ergänzung der Klassenarbeiten durch informelle oder formelle objektive Schulleistungstests (insbesondere bei Selektionsentscheidungen!)

Testtheoretische Kritik an mündlichen Prüfungen (1)

- Objektivität: Prüfung als Ausmaß der Beurteiler-Übereinstimmung (gleiche Prüfung, verschiedene Prüfer)
  - Durchführungsobjektivität: nicht immer gleiche Fragen und –abfolge, Reihenfolgen- und Kontrasteffekte
  - Auswertungs und Interpretationsobjektivität: oft mangelnde Kriterien für richtig und falsch, Ermittlung des Gesamtwertes/Gewichtung oft unklar
  - Birkel (1978): Objektivität zwischen 2 Beurteilern: .ca. .6
  - Höhere Objektivität wenn
    - sprachliche Leistung Beurteilungsgegenstand ist
    - Beurteiler geschult sind
    - Beurteilungs- und Gewichtungskriterien explizit sind

Testtheoretische Kritik an mündlichen Prüfungen (2)

- Reliabilität: Prüfung als Ausmaß der Beurteiler-Übereinstimmung,und zwar ...
  - gleiche Prüfung, gleiche Prüfer (= Retestreliabilität)
  - gleicher Prüfling, verschiedene Prüfer kurz hintereinander (= Paralleltestmethode), ca. .45)
  - Prüfer und Prüfling , ca. .50 oder besser

Testtheoretische Kritik an mündlichen Prüfungen (3)

- Validität: Prüfung durch Ermittlung von ...
  - Lehrplan/Lernziel-Repräsentanz der Fragen (Inhaltsvalidität)
  - Zusammenhängen mit anderen Kriterien wie Ergebnisse schriftlicher Prüfungen (.30), Noten, Berufserfolg (empirische Valdität: gleichzeitige oder prognostische Validität)
- Wirkung sachfremder Einflüsse (auf Seite der Situation, des Prüfers, des Prüflings) mindert Validität, z.B.
  - Sprechtempo (Birkel & Pritz, 1980), Primacy-, Recency-, Reihenfolgen-, Kontrasteffekte, Klassenzugehörigkeit, Brillenträger, Vorinformationen, Geschlecht (Prüfer und Prüfling), subjektive Maßstäbe, implizite Persönlichkeitstheorien usw.