(I) Wichtige Grundbegriffe der Testtheorie
© Herbert Paukert

Ein Test (Schularbeit) besteht aus verschiedenen Aufgaben (Items). Jede Aufgabe bezieht sich auf spezifische Leistungsfähigkeiten (Merkmale) eines Schülers. Jede Aufgabe wird mit Punkten bewertet, welche zwischen Null und einer maximalen Punkteanzahl liegen. Die bei der Aufgabenlösung erreichten Punkte eines Schülers spiegeln seine spezifische Leistungsfähigkeit wider. Dieses Ergebnis kann auch in Prozenten der bei der Aufgabe maximal erreichbaren Punkte ausgedrückt werden.

Jede Testaufgabe erfüllt bestimmte testtheoretische Anforderungen, die bei der Erstellung von Testaufgaben berücksichtigt werden sollen:

(1) Eindimensionalität. Eine Aufgabe soll sich auf ein oder nur wenige Leistungsmerkmale beziehen.

(2) Sprachliche Formulierung. Jede Aufgabe soll sprachlich klar, verständlich, einfach und eindeutig formuliert sein.

(3) Der Schwierigkeitsgrad einer Aufgabe. Er wird dadurch gemessen, dass für jede Aufgabe die Summe P der maximal erreichbaren Punkte und die Summe R der tatsächlich erreichten Punkte von allen Schülern einer Klasse gebildet werden. Dann wird der Prozentsatz von R bezogen auf P berechnet. Dieser Wert spiegelt die Schwierigkeit (eigentlich die Leichtigkeit) der Aufgabe in der vorliegenden Klasse von Schülern wider.

(4) Die Trennschärfe. Jede Testaufgabe erfüllt gesellschaftlich eine Selektionsfunktion, d.h. sie soll die „guten“ von den „schlechten“ Schülern trennen. Jene Schüler, welche das zu prüfende Leistungsmerkmal in hohem Ausmaß aufweisen, sind bei der Aufgabenlösung erfolgreich. Jene Schüler hingegen, welche das zu prüfende Merkmal nur in geringem Ausmaß aufweisen sind bei der Aufgabenlösung nicht erfolgreich.

Die Trennschärfe kann gemessen werden, indem man die Schüler in „gute“ (Hochgruppe) und in „schlechte“ Schüler (Tiefgruppe) einteilt. Dann ermittelt man in jeder Gruppe den Prozentsatz jener Schüler, welche die Aufgabe erfolgreich lösen. Die Differenz zwischen Hochgruppe und Tiefgruppe dient als Maß für die Trennschärfe. Sie sollte mindestens +10% betragen.

Offenkundig gibt es einen **Zusammenhang** zwischen Schwierigkeitsgrad und Trennschärfe. Sehr leichte Aufgaben werden von fast allen Schülern gelöst und sehr schwere Aufgaben von fast keinem Schüler. Also erreicht die Trennschärfe ein Maximum bei mittelschweren Aufgaben. Das Schaubild dieses funktionalen Zusammenhanges gleicht einem umgekehrten „U“. Bei der Zusammenstellung eines Tests ist es sinnvoll nur mittelschwere Aufgaben zu verwenden, weil diese am schärfsten trennen.

(5) Die Zuverlässigkeit (Reliabilität). Sie gibt an, wie genau ein Test das zu prüfende Merkmal misst. Man kann die Reliabilität bestimmen, indem man den Test in zwei Hälften mit gleichartigen Aufgaben zerlegt und diese den Schülern vorlegt. Die Bewertungsergebnisse in den beiden Testhälften sollten statistisch hoch miteinander korrelieren. (Split-Half-Methode). Eine andere Methode ist die Testwiederholung nach einiger Zeit. Dabei wird der originale Test mit seiner Wiederholung korreliert.

(6) Die Validität. Darunter versteht man, wie gut das Testergebnis mit einem Außenkriterium statistisch korreliert. Das wird auch als „Vorhersage-Gültigkeit“ bezeichnet.

(II) Die fünfstufige Notenskala

Wenn einer maximal erreichbaren Leistung genau 100% entsprechen, dann gibt es grundsätzlich zwei Varianten für die Grenze zwischen nicht genügender und genügender Leistung. Entweder liegt diese Grenze bei 50% oder bei 60% der maximal erreichbaren Leistung.

Note 5: Nicht genügende Leistung (falsch), von 0% bis 49% (oder von 0% bis 59%)
Note 4: Genügende Leistung (wenig richtig), von 50% bis 62% (oder von 60% bis 69%)
Note 3: Befriedigende Leistung (halb richtig), von 63% bis 75% (oder von 70% bis 79%)
Note 2: Gute Leistung (fast richtig), von 76% bis 88% (oder von 80% bis 89%)
Note 1: Sehr gute Leistung (richtig), von 89% bis 100% (oder von 90% bis 100%)

Hinweis: Für genügende Leistungen bei Fremdsprachen-Schularbeiten ab zehnter Schulstufe gilt, dass (1) mindestens 60% der maximalen Gesamtleisung erreicht werden müssen. Im so genannten Rezeptiv-Reproduktiv-Modell (RP) müssen (2) mindestens 50% der rezeptiven und (3) mindestens 50% der produktiven maximalen Gesamtleistung erreicht werden. Im Gesamtverrechnungs-Modell (GV) hingegen sind die Bedingungen (2) und (3) nicht verlangt, und die Teilaufgaben können auch verschieden gewichtet sein.

Hinweis: Die Mathematik-Schularbeiten ab zehnter Schulstufe müssen Typ-1-Aufgaben enthalten, die nur die Grundkompetenzen überprüfen und Typ-2-Aufgaben, in denen diese Grundkompetenzen angewendet und vernetzt werden. Für genügende Leistungen müssen die Typ-1-Aufgaben in überwiegendem Ausmaß (67%) richtig gelöst werden. Eine solche Aufgliederung erscheint mir übertrieben künstlich und wenig praktisch zu sein!

(III) Kriterienkataloge zur Beurteilung von Schularbeiten

(1) Wichtige Testformate zur standardisierten Erfassung von Leistungen:
Ja/Nein-Fragen, Multiple Choice, Lückentexte, Zuordnungsaufgaben, Diagramme

(2) Beurteilungskriterien in den sprachlichen Fächern:
Rezeptive Grundkompetenzen: (a) L, Lesen und (b) H, Hören
Produktive Grundkompetenzen: (c) SiK, Sprache im Kontext und (d) S, Schreiben

(3) Beurteilungskriterien der schriftlichen Textproduktionen (Schreiben):
(a) EA, Erfüllung der Aufgabenstellung (Inhalt)
(b) AL, Aufbau und Layout (Gliederung)
(c) SM, Spektrum sprachlicher Mittel (Ausdruck)
(d) SR, Sprach- und Schreibrichtigkeit (Grammatik)

(4) Beurteilungsrichtlinien in den naturwissenschaftlichen Fächern:
(a) TR, theoretische (gedankliche) Richtigkeit
(b) PR, praktische (sachliche bzw. rechnerische) Richtigkeit
(c) ORD, Ordnung, Gliederung und Übersichtlichkeit
(d) GEN, Genauigkeit der Ergebnisse

In sprachlichen Fächern werden die vier Kriterien in der Regel gleich gewichtet (25%). Die Kompetenz in jedem Kriterium wird mit charakteristischen, wohlunterscheidbaren Eigenschaften (Deskriptoren) beschrieben, wodurch für jedes Kriterium verschiedene Kompetenzstufen definiert werden. Das Erreichen einer Kompetenzstufe hängt vom Erfüllungsgrad der entsprechenden Deskriptoren ab (von falsch bis richtig). Zum Schluss müssen die erreichten Kriterienpunkte noch in Schulnoten umgerechnet werden. Diese Umrechnung wird im nächsten Kapitel (IV) detailliert beschrieben.

(IV) Ein formales Bewertungsverfahren von Tests und Schularbeiten

Ziel dieses Verfahrens ist die Objektivierung und Transparenz der Bewertung.
Das Excel-Arbeitsblatt „**noten.xls**“ von **Rosa Mistelbauer** und **Herbert Paukert** erlaubt die automatische Berechnung von Noten nach dem hier dargestellten Verfahren.

(1) Bestimmung der Anzahl N der Aufgaben (bzw. Kriterien) einer Arbeit.

(2) Bestimmung der maximal erreichbaren Punkte P_i einer Aufgabe (i):

Dabei gilt: $P_1 + P_2 + \ldots + P_N = P$ (maximale Punktesumme)

(3) Bestimmung der prozentuellen Gewichtung G_i einer Aufgabe (i):

Bei der Gewichtung wird für jede Aufgabe i ihre Gewichtszahl G_i bestimmt.
Sie spiegelt den prozentuellen Anteil der Aufgabe an der gesamten Arbeit wider.
Dabei gilt: $G_1 + G_2 + \ldots + G_N = 100$.
Die Gewichte entsprechen den verschiedenen Schwierigkeitsgraden der einzelnen Aufgaben innerhalb der Gesamtarbeit. Ihre Gesamtsumme muss immer 100% sein.
Will man die Gewichte den Maximalpunkten anpassen, dann gilt: $G_i = (100 / P) * P_i$

(4) Ermittlung des erreichten Punktewertes R_i für jede Aufgabe (i):

Nachdem jede Aufgabe (i) vom Lehrer sorgfältig korrigiert wurde, kann der erreichte Punktewert des Schülers in der Aufgabe bestimmt werden: $0 <= R_i <= P_i$.
Dabei gilt: $R_1 + R_2 + \ldots + R_N = R$ (erreichte Punktesumme $R <= P$)

(5) Ermittlung des gewichteten Punktewertes A_i für jede Aufgabe (i):

$\mathbf{A_i = R_i * (G_i / P_i)}$, mit $0 <= A_i <= G_i$

(6) Ermittlung des gewichteten Gesamtpunktewertes S für die ganze Arbeit:

$S = A_1 + A_2 + \ldots\ldots + A_N$, mit $0 <= S <= 100\%$

(7a) Ermittlung der Noten für die Arbeit (Fall a: Genügend ab 50%)

Die ***erste Rahmenbedingung*** ist, dass es 5 Noten gibt:
5 = nicht genügend, 4 = genügend, 3 = befriedigend, 2 = gut, 1 = sehr gut.
Die ***zweite Rahmenbedingung*** ist, dass ein „nicht genügend“ dann zu vergeben ist, wenn weniger als 50% des gewichteten Punktemaximums erreicht worden ist, also S < 50%.
Die ***dritte Rahmenbedingung*** ist, dass für die restlichen vier Noten, die sich von 50% bis 100% erstrecken, verbindliche Notengrenzen bestimmt werden.
Dabei sollten die Notenintervalle annähernd gleich breit sein.

Nicht genügend (5):	0	bis 49%	(49% breit)
Genügend (4):	50	bis 62%	(13% breit)
Befriedigend (3):	63	bis 75%	(13% breit)
Gut (2):	76	bis 88%	(13% breit)
Sehr gut (1):	89	bis 100%	(12% breit)

(7b) Ermittlung der Noten für die Arbeit (Fall b: Genügend ab 60%)

Die ***erste Rahmenbedingung*** ist, dass es 5 Noten gibt:
5 = nicht genügend, 4 = genügend, 3 = befriedigend, 2 = gut, 1 = sehr gut.
Die ***zweite Rahmenbedingung*** ist, dass ein „nicht genügend" dann zu vergeben ist, wenn weniger als 60% des gewichteten Punktemaximums erreicht worden ist, also S < 60%.
Die ***dritte Rahmenbedingung*** ist, dass für die restlichen vier Noten, die sich von 50% bis 100% erstrecken, verbindliche Notengrenzen bestimmt werden. Dabei sollten die Notenintervalle annähernd gleich breit sein.

Nicht genügend (5):	0 bis 59%	(59% breit)
Genügend (4):	60 bis 69%	(10% breit)
Befriedigend (3):	70 bis 79%	(10% breit)
Gut (2):	80 bis 89%	(10% breit)
Sehr gut (1):	90 bis 100%	(11% breit)

(8) Ein praktisches Beispiel (für Genügend ab 50%):

In unserem Beispiel soll die Testarbeit aus 7 Aufgaben mit folgenden Gewichten bestehen:

Aufgabe	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**…**	**…**	**…**	**Summe**
Punktemaximum	4	4	6	6	4	5	10				39
Aufgabengewicht	5	5	10	10	20	20	30				100 %

Die Arbeit eines Schülers wird nun vom Lehrer sorgfältig korrigiert und dann mit Hilfe der vorher angelegten Gewichtstabelle folgendermaßen bewertet:

Aufgabe	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**…**	**…**	**…**	**Summe**
Erreichte Punkte	4	2	6	3	2	2	6				25
Gewichtete Punkte	5	2.50	10	5	10	8	18				58.50
Gesamtnote											*Genügend*

Hinweis: Bei Aufsätzen in sprachlichen Fächern (z.B. „writing" in Englisch) wird der Aufsatz nach den vier Kriterien Inhalt (IT), Aufbau (AU), Ausdruck (AK), Grammatik (GK) beurteilt. Von den maximalen Punkten der Aufsatzbeurteilung entfallen jeweils ein Viertel auf jede dieser vier Kriterien. Wird beispielsweise der Aufsatz mit maximal 20 Punkten bewertet, dann erhält jedes Kriterium maximal 5 Punkte. Von den 100% der Gesamtarbeit kann der Aufsatz beispielsweise mit 30% bewertet werden. Dann haben die maximalen Punktewerte in den vier Beurteilungskriterien jeweils einen Prozentwert von 7.5%, und ein erreichter Kriterienpunkt entspricht dann dem gewichteten Prozentwert von 1.5%.
Vetoregel: Bei kompletter Themenverfehlung (IT = 0) wird der ganze Aufsatz mit 0 Punkten bewertet!

Schlussbemerkung: Der Vorteil der hier vorgeschlagenen formalen Bewertungsmethode ist, dass die Methode universell in allen Lehrfächern verwendet werden kann und damit zur Vereinheitlichung, Transparenz, Objektivierung und Vereinfachung der Notengebung beiträgt. Außerdem sollte jede Arbeit in eine Schülerstatistik und in eine Notenstatistik für das laufende Schuljahr eingetragen werden. In der Schülerstatistik wird für jeden Schüler seine Note eingetragen. In der Notenstatistik wird für jede Note ihre Häufigkeit eingetragen.