

Einem Vergleich verschiedener Klassifikationen schulpyschologischer Arbeitsfelder folgt die Darstellung der für die schulische Beratung wichtigen wissenschaftlichen Entwicklungen. Probleme, die sich aus der Anwendung von Tests für die schulische Beratung und Entscheidungsfindung ergeben, werden ebenso erörtert wie die Frage der Brauchbarkeit diagnostischer und therapeutischer Ansätze in der Beratung von Lehrer, Eltern und Schülern. Weitere Kapitel: einzelfallübergreifende Beratung, wünschenswerte Qualifikationen von Schulpsychologen, die Zusammenarbeit mit anderen Beratern.

Dr. Uwe Wiest ist Diplompsychologe und Leiter des Referates »Schulpsychologischer Dienst und Schullaufbahnberatung« beim Senator für Bildung Bremen.

Wiest Schulpsychologie

Uwe Wiest

Schul— —psychologie

Kohlhammer

ISBN 3-17-004893-7

wk

Das 1978 erschienene Buch ist in den entscheidenden Aussagen nach wie vor aktuell.

Insbesondere deswegen, weil heute Schulzensuren, Notendurchschnitte, für Schullaufbahnentscheidungen und Studienberechtigungen herangezogen werden, die metrisch unhaltbar und unfair sind.

Ein Beispiel: ich habe 1961 Abitur gemacht. Damals gingen in Schleswig-Holstein etwa 5 Prozent der Bevölkerung ins Gymnasium, der beste Schüler in meiner Klasse hatte einen Notendurchschnitt von 2,3, ich hatte einen von 2,7. Die Lehrkräfte nutzten die ganze Notenskala aus.

Mit meinem Zeugnis konnte ich studieren was ich wollte.

Wie kommt es, dass es heute viele Schüler gibt, die einen Abschluss mit einer 1 vor dem Komma haben, sogar 1,0? Sind das alle Goethes oder Einsteins?

Wie sieht es mit der Verteilung auf der Notenskala aus? eine 2,7 ist heute ja schon fast eine Beleidigung. Das heißt, die Schulnoten sind heute schlecht verteilt und differenzieren nicht mehr hinreichend. Als Maßstab für den Studienerfolg sagen sie sehr, sehr wenig.

Mehr dazu im folgenden Kapitel. Ich habe nur den Abschnitt über konkrete Tests weggelassen, weil dieser nicht mehr zeitgemäß ist.

1. Warum Tests?

Das Interesse der Schule an psychologischen Tests beruht hauptsächlich auf der Unzufriedenheit mit der pädagogischen Beurteilungspraxis. Die Überlegenheit von Tests wird dabei hinsichtlich ihrer inhaltlichen Aussage und ihrer Meßgenauigkeit erwartet.

1.1 Inhaltliche Konzepte, die psychologischen Tests zugrundeliegen

Psychologische Testverfahren lassen sich grob unterteilen in

- Intelligenztests, Begabungstests, Fähigkeitstests;
- Leistungstests;
- Persönlichkeitstests.

Ausgangspunkt für die Entwicklung von Tests der allgemeinen Begabung oder Intelligenz war die Beobachtung, daß die Lernfähigkeit von Kindern unterschiedlich ist (siehe Groffmann, 1963). Diese Lernfähigkeit oder »Intelligenz« ist relativ stabil, überdauernd, das heißt, nicht abhängig von situativen Bedingungen. Die Schulleistung, die sich in Zensuren kristallisiert, gibt dagegen nicht unbedingt darüber Auskunft, ob sich in ihr die Lernfähigkeit widerspiegelt oder auch Faktoren wie das Verhältnis zum Lehrer, augenblickliche Schwierigkeiten in der Familie, das Interesse des Schülers am Stoff, Ängste vor Prüfungssituationen.

Sowohl Binet (siehe Cronbach, 1960², S. 164) als auch Wechsler (siehe Priester, 1964, S. 229) betonen, daß »Intelligenz« eine Fähigkeit ist, die für die verschiedensten Aufgaben erforderlich ist, ganz gleich, ob es sich um sprachliche, numerische, motorische, zeichnerische Aufgaben handelt. Entsprechend verwenden beide Testautoren in ihren Verfahren sehr unterschiedliche Aufgaben, die untereinander hoch korrelieren. Das bedeutet: jemand, der bei dem einen Aufgabentyp gut abschneidet, wird vermutlich auch den anderen gut bewältigen können.

Bei Wechsler wird allerdings auch deutlich, daß er den einzelnen Aufgabentypen auch spezielle, von der globalen Intelligenz unabhängige Bedeutung zuschreibt. Daher hält er bei seinen Tests auch eine Interpretation der Untertestergebnisse für zulässig (hier muß

angemerkt werden, daß eine Profilinterpretation der Wechsler-Tests nicht aus inhaltlichen, wohl aber aus statistisch-metrischen Gründen problematisch ist). Wechslers Intelligenzkonzept entspricht der Zwei-Faktoren-Theorie von Spearman (siehe Hofstätter, 1972, S. 189f.), die besagt, daß es einen generellen Intelligenzfaktor (g-Faktor) und spezifische Aufgabenfaktoren (s-Faktoren) gibt (siehe Abbildung 1).

Der Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK, siehe Bondy, Hrsg., 1956) erlaubt die Berechnung eines Gesamt-Intelligenzquotienten (GIQ) sowie eines sprachlichen (VIQ) und Handlungs-Intelligenzquotienten (HIQ). Der Verbal- und der Handlungsteil bestehen aus je 5 Untertests, die im folgenden kurz beschrieben seien:

- | | |
|--|---|
| 1. Allgemeines Wissen (AW) | Beispiel: »Welches Tier gibt uns Milch?« |
| 2. Allgemeines Verständnis (AV) | »Warum werden Verbrecher eingesperrt?« |
| 3. Rechnerisches Denken (RD) | »Wieviel Stücke habe ich, wenn ich einen Apfel einmal durchschneide?« |
| 4. Gemeinsamkeiten finden (GF) | »Was ist das Gemeinsame bei Bier und Wein?« |
| 5. Wortschatztest (WT) | »Was ist ›laufen‹?« |
| Die Tests des Handlungsteils: | |
| 6. Zahlensymboltest (ZS) | Hier sollen Zeichen unter Zahlen geschrieben werden. Gewertet wird die richtige Zuordnung der Zeichen zu den Zahlen innerhalb von zwei Minuten. |
| 7. Bilderergänzen (BE) | Fehlende Details auf Bildern müssen erraten werden. |
| 8. Bilderordnen (BO) | Die Bilder von Bildgeschichten sollen in die folgerichtige Ordnung gebracht werden. |
| 9. Mosaiktest (MT) | Vorgegebene Muster sollen mit farbigen Würfeln nachgelegt werden. |
| 10. Figurenlegen (FL) | Zerschnittene gegenständliche Bilder sollen richtig zusammengelegt werden. |
| Zusatztest:
Zahlennachsprechen (ZN) | Zahlenreihen aus einstelligen Zahlen sollen nachgesprochen werden. |

Faktorenanalytische Verrechnungen von HAWIK-Daten ergeben meist um 5 voneinander unabhängige Faktoren, wobei man einen Faktor als Generalfaktor bezeichnen kann (siehe zum Beispiel Cohen, 1959).

Die Faktorenanalyse ist ein statistisches Verfahren, mit dem man die Interkorrelationen mehrerer Meßwertreihen in metrisch voneinander unabhängige (»orthogonale«) Faktoren umwandeln kann. Thurstone (1938, nach Cronbach, 1960²) fand aufgrund von 56 Tests, die er Studenten vorgelegt hatte, folgende Faktoren:

- | | |
|-------------------------------|--|
| V (Sprachverständnis) | Beispiel: »farbig« ist das gleiche wie:
groß <i>bunt</i> grau schön |
| W (Wortflüssigkeit) | »Schreibe so viele Wörter mit dem Anfangsbuchstaben ›S‹ auf wie Dir einfallen.« |
| N (Rechengewandtheit) | Addieren von Zahlenreihen. |
| S (Raumvorstellung) | Drehfiguren, Klappfiguren. |
| P (Wahrnehmungsschnelligkeit) | Fehlersuchen beim Vergleich zweier fast identischer Zahlenreihen oder Durchstreichen aller »e« in Buchstabenreihen. |
| M (assoziatives Gedächtnis) | Einprägen von Zahlen-Wörter-Verbindungen (Stuhl-21, Haus-45), Einsetzen der richtigen Zahlen hinter die Wörter auf einem Antwortbogen. |
| R (schlußfolgerndes Denken) | a b c b c d c d e f e f
Durchstreichen des Buchstabens, der nicht in die Zeile paßt. |

Nach Thurstones Intelligenztheorie (Multiple-Faktoren-Theorie) sind diese Faktoren mit unterschiedlichem Gewicht in den verschiedenen Denkleistungen enthalten (siehe Abbildung 1). Faktorenanalysen vor allem von Guilford (siehe Ulmann, 1968, S. 45f.) ergaben inzwischen weit mehr und differenziertere Faktoren. Bemerkenswert ist, daß Guilford in seinem System der intellektuellen Fähigkeiten auch das kreative Denken nennt.

Kreative Aufgaben unterscheiden sich von den Aufgaben üblicher Intelligenztests dadurch, daß nicht eine bestimmte richtige Lösung gefunden werden soll (convergentes Denken), sondern möglichst viele Antworten (Flüssigkeit), Antworten, die man in möglichst viele verschiedene Kategorien einordnen kann (Flexibilität) oder

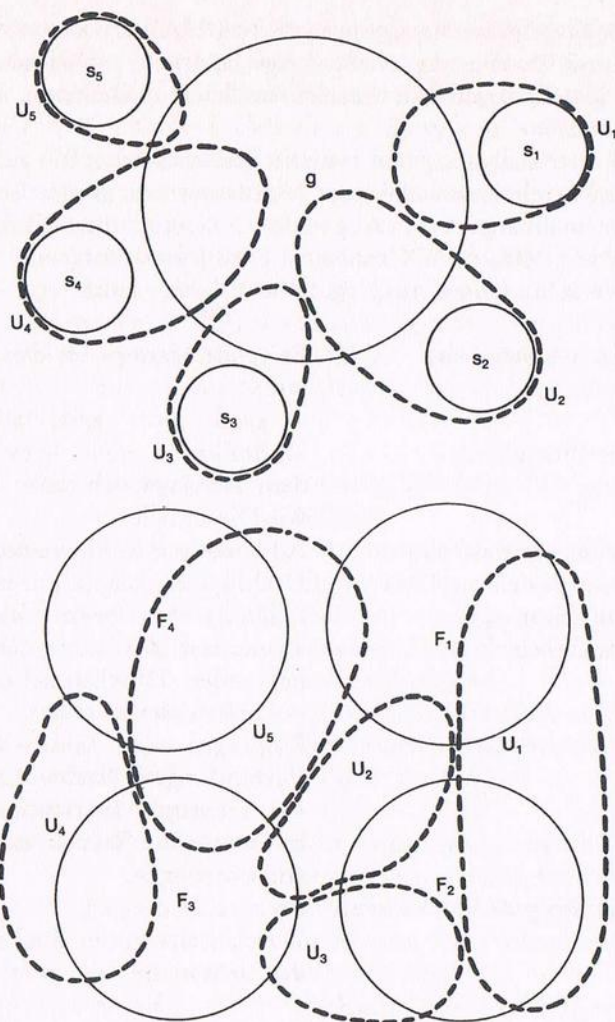


Abbildung 1: oben: Zwei-Faktoren-Theorie von Spearman. g = Generalfaktor; $s_1 - s_5$ = spezifische Faktoren; $U_1 - U_5$ = Untertests. unten: Multiple-Faktoren-Theorie von Thurstone. $F_1 - F_4$ = Faktoren; $U_1 - U_5$ = Untertests.

Antworten, die besonders selten vorkommen (Originalität). Daher nennt man Kreativität auch divergentes Denken. Ulmann (1968) konnte anhand der Literatur aufzeigen, daß der Zusammenhang zwischen kreativem Denken und Schulleistungsangaben längst nicht so deutlich ist wie zwischen Intelligenz und Schulleistung. Eine Übersicht über die Methode der Faktorenanalyse gibt Lienert (1969³).

Beide, nämlich die Zwei- und die Multiple-Faktoren-Theorie, haben ihre Berechtigung. Es ist nämlich außerordentlich schwer, faktorenpure Tests zu konstruieren, das heißt, Tests, die untereinander nicht korrelieren. Daher lassen sich bei Testbatterien, die die Faktoren Thurstones wiedergeben sollen, sowohl die beabsichtigten Faktoren als auch ein allgemeiner Begabungsfaktor nachweisen (siehe Horn, 1962).

Intelligenz oder Begabung, wie die Tests sie messen, sind nun Konstrukte, die die Eigenschaft haben, sich im Laufe der Entwicklung des Kindes zu verändern. Cohen (1959) wies für die amerikanische Originalform des HAWIK (der WISC) nach, daß die Intelligenz von 7-, 10- und 13jährigen Kindern eine recht unterschiedliche Faktorenstruktur aufweist. Hofstätter (siehe Hofstätter, 1972) setzte die Intelligenzleistungen von Kindern verschiedenen Alters miteinander in Beziehung und fand drei Faktoren:

F_1 (Säuglingsalter): senso-motorische Wachheit;

F_2 (2–4 Jahre): Persistenz, Durchhaltefähigkeit bei einmal begonnenen Beschäftigungen;

F_3 (über 4 Jahre): Symbolverständnis und sprachliches Denken.

In diesem Zusammenhang ist auch auf die Entwicklungstheorie von Piaget (Piaget und Inhelder, 1972) hinzuweisen. Die entsprechenden Forschungsarbeiten machen deutlich, daß die Intelligenzleistungen verschieden alter Kinder qualitativ unterschiedlich sind. Begabungstestergebnisse sind immer auf die Altersgruppe bezogen, das heißt, die absolute Leistung eines Kindes in einem Test wird mit den Leistungen von Kindern seines Alters verglichen. Das ist deswegen notwendig, weil die absolute Leistung von Kindern mit zunehmendem Lebensalter ansteigt. Vergleicht man die durchschnittliche Testleistung verschiedener Altersgruppen, so erweisen sich die 20–25jährigen als am leistungsstärksten, danach sinken die absoluten Testwerte ab, und zwar wenig bei sprachlichen, stärker dagegen bei Handlungsaufgaben. Das bedeutet aber nicht, daß die Intelligenz identischer Personen tatsächlich mit zunehmendem Alter abnimmt – die Testnormen enthalten ja Werte verschiedener Generationen, die zu demselben Datum erhoben wurden. Längsschnittuntersuchungen haben gezeigt, daß die Intelligenz bis zum fünfzigsten Lebensjahr relativ konstant bleibt (siehe Hofstätter, 1972).

1.2 Die Meßeigenschaften von Tests

Die Meßgenauigkeit eines psychologischen Tests ergibt sich aus der Art seiner Konstruktion.

Zu Beginn der Testerstellung definiert man das Merkmal, das gemessen werden soll, und überlegt sich, welche Testaufgaben geeignet sind, dieses Merkmal zu prüfen. Neben der inhaltlichen Frage ist auch zu entscheiden, welchen Aufgabentyp man wählt, zum Beispiel gebundene Aufgaben, bei denen der Proband ankreuzt, oder freie Aufgaben, die etwa durch ein Wort oder einen Satz beantwortet werden.

Sodann werden die Aufgaben einer Analyse unterzogen. Sie werden einer Stichprobe von Probanden vorgelegt. Anhand der Lösungen wird eine Bestimmung der Aufgabenschwierigkeiten vorgenommen: die Schwierigkeit einer Aufgabe berechnet man aus dem Quotienten zwischen der Anzahl der Personen, die die Aufgabe richtig gelöst haben, durch die Gesamtzahl der Personen. Bei Niveautests, also Tests ohne oder mit einer großzügigen Zeitbegrenzung, verwendet man sowohl Aufgaben mit geringer, mittlerer als auch hoher Schwierigkeit.

Bei der Trennschärfeanalyse prüft man, ob die Aufgabenlösungen möglichst hoch mit der Gesamtestleistung korreliert sind – ob die Aufgabe also möglichst trennscharf zwischen »guten« und »schwachen« Testprobanden unterscheidet. Je höher die Trennschärfe der Aufgaben ist, desto eindimensionaler (faktorenlerner) ist der Test. Nach der Aufgabenanalyse wird die Testendform hergestellt. Dann erfolgt die Testeichung: der Test wird einer möglichst repräsentativen Stichprobe von Probanden vorgelegt. Durch Transformation der Rohwerte (zum Beispiel: Anzahl richtiger Lösungen) in Standardmeßwerte der Normalverteilung gewinnt man Testnormen. Normen können für die Gesamtstichprobe, aber auch für Teilstichproben (Altersnormen, Geschlechts- und Sozialstatusnormen zum Beispiel) bestimmt werden.

Die Gütekriterien des Tests werden anhand der Analysestichprobe, besonders aber nach der Eichung ermittelt. Die Gütekriterien für psychologische Tests sind:

Objektivität: Ein Test ist objektiv, wenn verschiedene Auswerter des Tests zu demselben Ergebnis kommen. Tests, die mit Schablonen oder gar einer Datenverarbeitungsanlage ausgewertet werden, haben die höchste Objektivität. Die Objektivität sinkt, wenn dem Auswerter individueller Spielraum der Interpretation bleibt.

Zuverlässigkeit (Reliabilität): Ein Test ist zuverlässig, wenn er bei einer Testwiederholung das gleiche Ergebnis liefert (Re-Test-Zuverlässigkeit) oder wenn ein Paralleltest (ein inhaltsgleicher Test mit anderen Aufgaben) das gleiche Resultat ergibt (Paralleltestzuverlässigkeit). Neben der Testwiederholungsmethode und dem Vergleich zweier Parallelformen eines Tests gibt es noch die Test-

halbierungsmethode – das heißt, etwa alle Aufgaben mit gerader Zahl in der Reihenfolge werden mit allen Aufgaben mit ungerader Zahl korreliert. Ein spezielles und häufig angewandtes Verfahren ist die Konsistenzanalyse: dabei geht man davon aus, daß man einen Test nicht nur in zwei Hälften, sondern in so viele Teile untergliedern kann wie Aufgaben vorhanden sind.

Gültigkeit (Validität): Ein Test ist dann gültig, wenn er das zu prüfende Merkmal wirklich erfaßt (und nicht ein ähnliches oder gar ein anderes Merkmal) und wenn der Test zuverlässig ist.

Zuverlässigkeits- und Gültigkeitsangaben geschehen in der Form von Korrelationskoeffizienten. Die Zuverlässigkeit sollte den Wert von $r_{tt} = 0,9$ möglichst nicht unterschreiten (t steht für »Test« oder »Testteil«). Wenn Gültigkeitskoeffizienten die Größenordnung von $r_{tc} = 0,5$ oder $0,6$ erreichen, (c steht für das Kriterium, also das zu testende Merkmal), ist man schon recht zufrieden. Ein r_{tc} von $0,5$ besagt, daß das Testergebnis $0,5^2$, also ein Viertel der Testleistung erfaßt.

Gültigkeitskoeffizienten sind oft relativ niedrig, weil das Kriterium, zum Beispiel die Schulzensur, nicht sehr zuverlässig ist. Manchmal ist auch kein geeignetes quantitatives Kriterium vorhanden, so daß gar keine Gültigkeitskoeffizienten bestimmt werden können.

Man unterscheidet folgende Gültigkeitsarten:

- Die Vorhersagegültigkeit (»predictive validity«): hier wird das Testergebnis mit einem Kriterium verglichen, dessen Merkmalsausprägung erst später erfolgt. So korreliert der »Fremdsprachen-Eignungstest für die Unterstufe« (FTU 4–6) von Correll und Ingenkamp $r_{tc} = 0,56$ mit den Englischnoten nach zwei Jahren.
- Die Übereinstimmungsgültigkeit (»concurrent validity«): hier vergleicht man das Testergebnis mit einem gleichzeitig erhobenen Kriterium. Ein Beispiel ist die Korrelation zwischen dem Ergebnis einer Aufnahmeprüfung für das Gymnasium und einem im gleichen Zeitraum erhobenen Intelligenztest. In beiden genannten Beispielen würde man auch von »äußerer Gültigkeit« sprechen, da die Testergebnisse mit anderen Kriterien verglichen wurden. Die Validität von Intelligenztests wird dagegen oft durch Korrelation mit einem bereits anerkannten Test bestimmt (was Boring bereits 1923 zu der ironischen Definition »Intelligenz ist das, was die Intelligenztests messen« veranlaßte; siehe Hofstätter, 1972). In diesem Fall spricht man von »innerer Gültigkeit«.

- Die inhaltliche Gültigkeit (»content validity«): bei dieser Art der Gültigkeit wird kein Koeffizient berechnet. Vielmehr prüft man, ob die Aufgaben »repräsentativ« (Michel, 1964) für den zu testenden Inhalt sind. Ein Rechtschreibtest, bei dem Wörter geschrieben werden müssen, die aus den drei am häufigsten benutzten Rechtschreibwerken der betreffenden Klassenstufe ausgewählt wurden, ist inhaltlich gültig. Lienert (1969³) schlägt vor, im Zweifelsfall Experten die inhaltliche Gültigkeit einschätzen zu lassen.

Bei ganz offensichtlicher Gültigkeit spricht man auch von logischer oder trivialer Gültigkeit. Diese ist deutlich abzugrenzen von der Schein-Gültigkeit (»face validity«), bei der Testaufgaben dem Laien so erscheinen, als messen sie das angegebene Merkmal, obwohl der Test einer wissenschaftlichen Überprüfung nicht standhält.

- Die Konstrukt-Gültigkeit (»construct validity«): soll ein Test kein bestimmtes Verhaltensmerkmal überprüfen, sondern ein Konzept, das nicht so ohne weiteres operationalisierbar ist (zum Beispiel »Angst«, »Soziale Anpassung«, »Aggressivität«), kann man so vorgehen, daß man aus einer Persönlichkeitstheorie heraus Hypothesen bildet – etwa, indem man vorhersagt, wie sich »sozial angepaßte« und »sozial unangepaßte« Individuen nach der betreffenden Theorie in manchen Situationen unterschiedlich verhalten müßten. Lassen sich solche Verhaltensdifferenzen zwischen der Gruppe mit hohem Testwert in »Angepaßtheit« und der Gruppe mit niedrigem Testwert nachweisen, die der Theorie entsprechen, kann man mit einer gewissen Berechtigung von der (Konstrukt-)Validität des Tests ausgehen.

Überprüft man die Faktorenstruktur eines mehrdimensional angelegten Tests mittels der Faktorenanalyse, so kann man im Falle des Erfolges ebenfalls von »Konstrukt-Gültigkeit« sprechen.

Die Meßgenauigkeit von Tests ist, was Objektivität und Zuverlässigkeit angeht, also relativ leicht zu beurteilen. Im allgemeinen schneiden allgemeine und spezielle Fähigkeitstests hinsichtlich dieser beiden Kriterien auch brauchbar ab. Der entscheidende Aspekt aber, die Gültigkeit, ist schwer zu beurteilen, und selbst ein Koeffizient von beachtlicher Höhe besagt nicht unbedingt, daß der Test für eine bestimmte Fragestellung brauchbar ist. Das soll im folgenden an schulischen Entscheidungen aufgezeigt werden:

1.3 Veralten von Tests

Sowohl Testnormen als auch die numerisch ermittelte Gültigkeit von Tests verlieren ihre Aussagefähigkeit, und zwar wegen der Veränderung des Kriteriums.

Ermittelt man den Standardwert eines Testprobanden, der 11 Jahre alt ist, vergleicht man seine Leistung im Grunde nicht mit den Testleistungen seiner Altersgruppe, sondern mit der Gruppe der 11jährigen Kinder des Jahres 1969 (Prüfsystem für Schul- und Bildungsberatung von Horn) oder gar 1954 (HAWIK). Bei der Anwendung dieser Testverfahren setzt man also im Grunde voraus, daß die Veränderungen in Schule und Gesellschaft (eventuell auch biologische Veränderungen durch eine andere Ernährung und gesundheitliche Fürsorge) auf die Kinder und Jugendlichen völlig ohne Einfluß geblieben sind. Tatsächlich haben die Testleistungen sich aber verändert, wie Küffner und Rothe (1976) für Begabungs-, Leistungstests und Persönlichkeitsfragebögen nachweisen konnten. Royl und Schwarzer (1976) stellten für das Prüfsystem für Schul- und Bildungsberatung (PSB), bezogen auf Kinder im 5. Schuljahr, eine erhebliche Leistungssteigerung gegenüber der Normstichprobe fest. Verändert sich die absolute Leistung, dann ist auch wahrscheinlich, daß sich die Interkorrelationen und die Faktorenstruktur von Tests verändern. Ebenso ist anzunehmen, daß Gültigkeitsindizes nicht konstant bleiben, etwa weil die Lehrpläne der Schulen sich verändert haben. (So kann man aus den Bremer Lernplänen entnehmen, daß die Rechtschreibleistung in der Orientierungsstufe im Rahmen des Deutschunterrichts nur noch ein Fünftel der schriftlichen Note ausmachen soll; Der Senator für Bildung, 1977.) Dies gilt stärker für Leistungstests als für Begabungstests.

Sowohl Küffner und Rothe als auch Royl und Schwarzer fordern, daß Tests in kurzen Zeitabständen neu normiert werden müssen, andernfalls nehmen die unzureichenden Beratungen in der Schule im Laufe der Zeit zwangsläufig zu.

1.4 Testgültigkeit und Qualität personeller Entscheidungen

Wenn eine Aussage über das Ausmaß der Gültigkeit eines Tests vorliegt, wissen wir damit noch immer nicht viel über die Brauchbarkeit des Verfahrens für eine spezielle Entscheidung. Das gilt auch für Entscheidungen während der Schullaufbahn. Der Grund ist darin zu sehen, daß es bei Schullaufbahnentscheidungen um Selektions- oder Klassifikationsentscheidungen geht (siehe Lang-

feldt-Nagel und Langfeldt, 1976; Tröger und Lüking, 1977; Bachmair, 1976).

Bei der Selektion oder Auslese ist zu entscheiden, ob ein Schüler für einen Ausbildungsgang angenommen oder abgelehnt wird. Zwei Alternativen stehen zur Auswahl.

Bei der Klassifikation geht es darum, den Schüler verschiedenen Ausbildungsgängen A, B, C; . . . N zuzuordnen. Metrisch stellt die Aufgabe der Klassifikation also eine Erweiterung der Auslesefragestellung dar.

Gehen wir im einfachsten Fall von einer Auslesefragestellung aus: wir haben einen Test, der einen Standardwert liefert, den wir als Maß der »Begabung« ansehen wollen. Wir kennen die Validität des Tests und wollen versuchen, für einen Bildungsgang so viele geeignete Schüler wie möglich auszuwählen. Außerdem ist die Anzahl der Bewerber für das Bildungsangebot und die Zahl der vorhandenen Plätze bekannt. Und schließlich wollen wir annehmen, daß die bisherige Auswahl nach dem Zufall erfolgt ist, daß also bisher kein valides Ausleseinstrument vorgelegen hat (was, wie noch aufzuzeigen ist, eine Annahme darstellt, die für schulische Entscheidungen nicht zutrifft).

Auf der Abszisse (Abbildung 2) sind die Werte des Tests aufgetragen, auf der Ordinate die des Kriteriums (»Schulerfolg«). Die Schüler haben in diesem Koordinatensystem eine Position, die sich aus dem Testwert X und dem Kriteriumswert Y ergibt. Alle Schülerpositionen befinden sich innerhalb der Ellipse.

Die Form der Ellipse variiert mit der Vorhersagegültigkeit des Testverfahrens: ist $r_{tc} = 1$, wird sie zu einer Geraden, je höher die Gültigkeit, desto langgestreckter wird die Ellipse sein. Besteht kein Zusammenhang zwischen Test und Kriterium, verteilen sich die Schülerpositionen wahllos über das gesamte Feld.

Bei dem Beispiel in Abbildung 2 gehen wir davon aus, daß sich unter den Bewerbern 50 Prozent Geeignete befinden. Schüler unterhalb der Position Y_i werden sich also nicht bewähren, die Schüler oberhalb werden erfolgreich sein, wenn man sie zuläßt.

Da sehr viele Bewerber vorhanden sind, setzen wir als Aufnahmekriterium einen überdurchschnittlich hohen Testwert fest, wir werden also die Mehrzahl der Bewerber ablehnen (X_i). Ein hoher kritischer Testwert entspricht einer niedrigen Selektionsquote – denn wir wählen im Vergleich zu der Bewerberzahl nur einen kleinen Teil aus.

Die beiden Geraden, die von X_i und Y_i ausgehen, teilen die Ellipse, die Schülerverteilung, in vier Gruppen:

Auf der Seite der akzeptierten Schüler sind es zwei Gruppen: A,

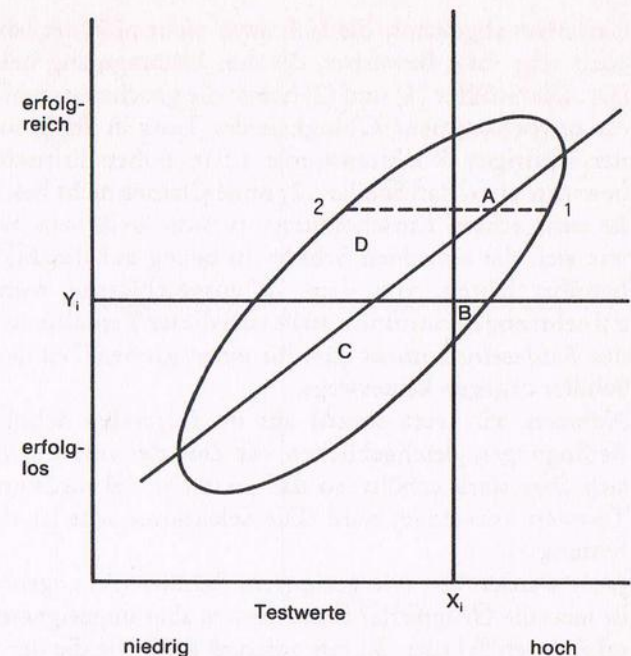


Abbildung 2: X_i : kritischer Testwert – unterhalb dieses Wertes werden die Schüler nicht angenommen, oberhalb angenommen (der Wert wird auch »cutoff point« genannt). Ein hoher kritischer Testwert entspricht einer niedrigen Selektionsquote.

Y_i : kritischer Kriteriumswert. Er besagt, wieviele erfolgreiche Schüler in der vorigen Bewerbergruppe enthalten waren. In diesem Falle sind es 50 Prozent. Y_i liegt hoch, wenn die Zahl erfolgreicher Schüler (die Grundquote) klein ist.

A: zu Recht Angenommene (angenommen und erfolgreich)

B: zu Unrecht Angenommene (angenommen und erfolglos)

C: zu Recht Abgelehnte (abgelehnt und erfolglos)

D: zu Unrecht Abgelehnte (abgelehnt und erfolgreich).

1 und 2: zwei Schüler, die beide befähigt sind, 1 wurde aufgrund des Tests empfohlen, 2 abgelehnt.

die Gruppe der Schüler, die später auch erfolgreich sein wird, und B, die Gruppe der Schüler, die ausgewählt wurde, die aber später mit dem Bildungsangebot nicht zurechtkommt. Diese Gruppe ist in unserem Beispiel verschwindend klein und spielt im Vergleich zu A kaum eine Rolle. Der Test hat sich als Auslesekriterium also bewährt, obwohl die Form der Ellipse anzeigt, daß die Validität des Tests nur eine begrenzte Höhe aufweist.

Er hat sich aber nur für die Gruppe der angenommenen Schüler bewährt, wie die Seite links von X_i zeigt: zwar wurden viele

Bewerber abgelehnt, die sich auch nicht bewährt hätten (C), aber auch sehr viele Bewerber, die den Bildungsgang bewältigt hätten (D). Die Schüler (1) und (2) haben die gleiche tatsächliche Eignung, die unvollkommene Gültigkeit des Tests in der Kombination mit der niedrigen Selektionsquote (dem hohen kritischen Testwert) bewirkte aber, daß Schüler (2) seine Chance nicht bekam.

In einer echten Entscheidungssituation weiß man natürlich nicht, wie sich die einzelnen Schüler in bezug auf das Bildungsangebot bewährt hätten, von dem sie ausgeschlossen wurden. Für die aufnehmende Institution stellt sich dieser Test also als hervorragendes Ausleseinstrument dar, für einen großen Teil der abgelehnten Schüler dagegen keineswegs.

Nehmen wir jetzt einmal an, im folgenden Schuljahr sind alle Bedingungen gleichgeblieben, die Zahl der vorhandenen Plätze hat sich aber stark erhöht, so daß jetzt ein viel niedrigerer kritischer Testwert verwendet wird. Die Selektionsquote ist also hoch (Abbildung 3).

Jetzt werden fast alle geeigneten Schüler (A) angenommen, dafür ist jetzt die Gruppe der akzeptierten, aber ungeeigneten, weil später erfolglosen Schüler (B) fast genauso groß wie die der später erfolgreichen (A). Auf der anderen Seite sind die abgelehnten Schüler (C und D) fast alle wirklich ungeeignet (C).

Die Zuweisung aufgrund des Tests ist für die Institution, das Bildungsangebot, eine Katastrophe – auf der anderen Seite können die Schüler, die bei einer niedrigen Selektionsquote aufgrund des Tests irrtümlich abgelehnt worden wären, den Bildungsgang erfolgreich durchlaufen.

Der Vergleich beider Entscheidungssituationen zeigt noch etwas: wenn eine aufnehmende Institution wünscht, daß sie nur Schüler bekommt, die später auch erfolgreich sind, und wenn sie kein völlig valides Ausleseinstrument besitzt, verwendet sie eine so niedrige Selektionsquote, daß die Gruppe B nicht mehr vorhanden ist. Dieser Punkt wird weiter rechts auf der Abszisse liegen, wenn die Ellipse eine große Nebenachse hat, ist die Ellipse dagegen schlanker (die Testvalidität also höher), kann der kritische Testwert weiter links liegen.

Aus der Betriebspsychologie liegen Wahrscheinlichkeitstafeln vor, die eine Vorhersage der Zahl der Erfolgreichen erlaubt, wenn man die Gültigkeit des Tests, die Selektionsquote und die Zahl der bisher Erfolgreichen innerhalb ihrer Gruppe kennt (Taylor und Russell, 1939, siehe McCormick und Tiffin, 1974).

Diese Tafeln zeigen, daß ein wenig valider Test die Auslese noch erheblich verbessert, wenn man eine sehr niedrige Selektionsquote

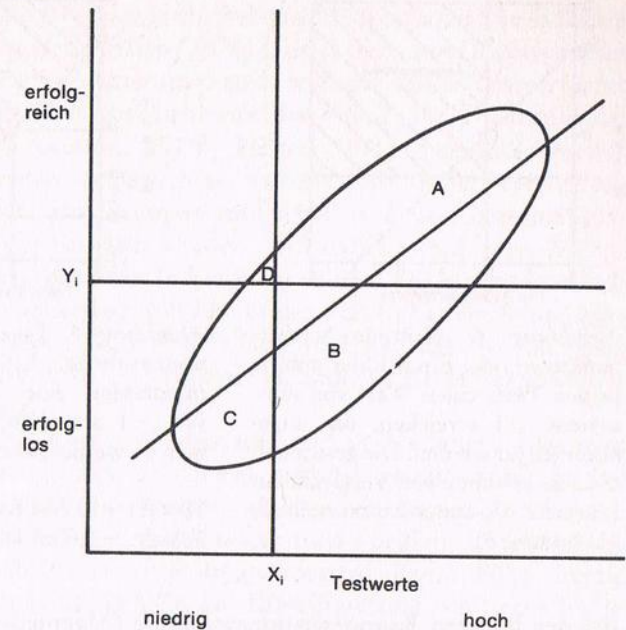


Abbildung 3: Entscheidung bei hoher Selektionsquote (Erklärungen siehe Abbildung 2).

verwendet. Bei einer sehr hohen Selektionsquote, wenn man also praktisch alle Bewerber akzeptiert, ist selbst der valideste Test wertlos.

1.5 Kombinierte Testwerte

Reicht die Gültigkeit einzelner Testverfahren nicht aus, um das Erfolgskriterium hinreichend genau vorherzusagen, kann man verschiedene Verfahren kombinieren.

Bei der multiplen Schnittpunktmethode legt man einfach für jeden verwendeten Test getrennt den kritischen Wert fest. Angenommen werden diejenigen Bewerber, die in allen Verfahren oberhalb der kritischen Werte liegen. Das bedeutet also, daß man eine zu schwache Einzeltestleistung nicht durch eine gute Leistung in einem anderen Test ausgleichen kann (Abbildung 4). Nach diesem Modell wird in der Schule vorgegangen, wenn man eine ungenügende Leistung in Mathematik nicht durch eine sehr gute Leistung in Deutsch ausgleichen kann und die Klasse wiederholen beziehungsweise die Schulart verlassen muß.

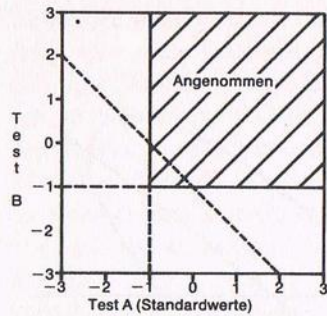


Abbildung 4: Multiple Schnittpunktmethode. Ein Schüler muß in beiden Tests einen Wert von mindestens -1 erreichen, um angenommen zu werden. Die gestrichelte Linie erlaubt einen Vergleich zur Linearen Kompensationsmethode (Abbildung 5).

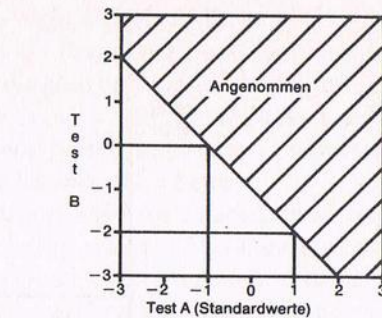


Abbildung 5: Lineare Kompensationsmethode. Ein Schüler muß mindestens eine Testwertsumme von -1 erreichen, um angenommen zu werden, also zum Beispiel:

Test A: +1, Test B: -2;

Test A: -1, Test B: 0.

Bei der linearen Kompensationsmethode (Magnusson, 1975²) legt man eine für die Testverfahren gemeinsame Selektionsquote fest. Ob ein Proband über dem gemeinsamen kritischen Wert liegt, ergibt sich aus der einfachen Addition seiner Testwerte (Abbildung 5). Dieses grobe Verfahren kann dadurch verfeinert werden, indem man mittels einer multiplen Regressionsanalyse (Lienert, 1969³, Magnusson, 1975²) die einzelnen Testverfahren so gewichtet, daß sich die für die Testbatterie größtmögliche Gesamtvalidität ergibt.

1.6 Klassifikation

In den bisherigen Überlegungen sind wir von Entscheidungen des Typus »Annehmen/Ablehnen« ausgegangen. Der Auslesegesichtspunkt ist allerdings für viele schulische Entscheidungen unangemessen, zum Beispiel, wenn es darum geht, Schüler zu mehreren Schularten zuzuordnen (Hauptschule, Realschule, Gymnasium) oder für Differenzierungsgruppen aufzuteilen.

Die Klassifikationsaufgabe ist komplexer, weil mehrere Alternativen vorhanden sind, und sie erlaubt es nicht, die »Abgelehnten« einfach zu vergessen, wie das bei betriebspsychologischen Entscheidungen (Einstellungen) oft der Fall ist.

Um Schüler verschiedenen Bildungsangeboten zuzuordnen, könnte man sich der Tests bedienen, die am besten zwischen Schülern dieser Bildungsangebote unterscheiden. Andererseits

könnte man die Gruppen (zum Beispiel Hauptschüler, Realschüler, Gymnasiasten, siehe Heller, 1976) hinsichtlich ihrer Eigenschaften deutlicher trennen, wenn man auch weniger valide Testverfahren mit einbezieht. Das geschieht mit der »multiplen Diskriminanzfunktion« (Magnusson, 1975², Heller, 1976). Die individuellen Testwerte werden so gewichtet, daß die Differenzen der Mittelwerte zwischen den Gruppen möglichst groß und innerhalb der Gruppen möglichst klein werden. Anhand der multiplen Diskriminanzfunktion kann ein Individuum sodann der für ihn passendsten Gruppe zugeordnet werden. Heller (1976) hat die Brauchbarkeit automatischer (Computer-)Klassifikation mittels der multiplen Diskriminanzanalyse im Vergleich zu anderen Klassifikationsstrategien aufgezeigt.

Ein zukunftssträchtiges nichtparametrisches Verfahren (das also nicht voraussetzt, daß die Daten normal verteilt sind) ist die »Konfigurationsfrequenzanalyse« (KFA) von Krauth und Lienert (1973). Mittels dieses Verfahrens kann man prüfen, ob bestimmte Merkmalskombinationen (Konfigurationen) überzufällig häufig auftreten. Man kann die KFA zur Klassifizierung von Lernschwierigkeiten einsetzen und so Gesichtspunkte für eine differentielle Förderung gewinnen (siehe Masendorf und Roeder, 1974).

1.7 Übereinstimmungs- und Vorhersagegültigkeit psychologischer Tests im schulischen Raum, Vergleich mit schulischen Beurteilungen

Gebauer (1965) verwendete mehrere Intelligenztests sowie die Ergebnisse von Aufnahmeprüfungen, um diese Prediktoren mit dem späteren Schulerfolg im Gymnasium zu vergleichen. Er berechnete biseriale Korrelationen (Beziehung zwischen den normalverteilten und stetigen Testwerten beziehungsweise Noten und dem Kriterium »im Gymnasium«/»nicht im Gymnasium«, einem dichotomischen Merkmal).

Die Korrelationen fielen, je nach Testverfahren und Ort der Erhebung, recht unterschiedlich aus. Die Tests korrelierten mit dem Schulerfolg nach einem halben Jahr zwischen $r_{tc} = .44$ bis $.62$, nach sechs Jahren zwischen $.45$ und $.65$.

Die Beziehung zwischen den Aufnahmeprüfungen und dem späteren Schulerfolg: nach einem halben Jahr: $.14$ bis $.31$, (in einem Fall) nach einem Jahr $.02$, (in einem Fall) $.45$ nach 6 Jahren. Tent (1969) ermittelte zwischen dem Ergebnis des Leistungsprüfsystems von Horn am Ende des vierten Schuljahres und der Schulartzugehörigkeit im 5. Schuljahr (triseriale Korrelation) einen Wert von $.59$.

Außerdem kommt Tent für einen Zeitraum von drei Schuljahren auf eine Schätzung der Vorhersagegültigkeit von etwa .70. Tent versucht zu begründen, daß eine höhere Testgültigkeit nicht erreichbar ist, weil das Kriterium, die schulische Beurteilung, nicht zuverlässig genug ist.

Heller (1976) fand Korrelationen zwischen Schülerklassifikationen mit der multiplen Diskriminanzanalyse (automatisches Klassifikationsverfahren, AUKL) und dem Lehrerurteil von .49 bis .59. Außerdem kommt Heller (1977) zu dem Ergebnis, daß Intelligenztests allein nur noch für die Prognose des Schulerfolgs geeignet sind, wenn eine hohe Intelligenz vorliegt.

Kastner (1978) ermittelte für die »Aufgaben zum Nachdenken (AzN4)« und Lehrerurteilen hinsichtlich der Eignung für weiterführende Schularten eine Korrelation von .77 (60% des Lehrerurteils wurde damit erklärt). Das Prüfsystem für Schul- und Bildungsberatung (PSB) erklärte 35% des Lehrerurteils ($r_{tc} = .60$). Beide Tests zeigten außerdem unterschiedlich hohe Beziehungen zu Schulnoten. Bedenklich niedrig war die Korrelation beider Tests miteinander (innere Gültigkeit, siehe Abschnitt 1.2 dieses Kapitels, Seite 46): sie betrug nur .70 – damit sind die Tests nicht einfach austauschbar. Der Verfasser führt die verhältnismäßig niedrige Beziehung beider Tests und die geringere Korrelation des PSB mit der Schulleistung darauf zurück, daß der PSB ein mehrfaktorieller Test mit Sinne Thurstones ist (siehe Abschnitt 1.1 dieses Kapitels, Seite 43).

1.8 Nutzen durch Tests und Kosten

Cronbach und Gleser (1965²) haben theoretische Überlegungen und Hypothesen beigetragen, in welchem Verhältnis der Nutzen durch angewendete Tests zu den entstehenden Kosten steht.

Die im Abschnitt 1.4 dargestellte Entscheidungstheorie muß dahingehend relativiert werden, daß die Verbesserung einer Entscheidung durch Tests nicht an einem Zufallsergebnis gemessen werden kann, sondern an der Validität des vorhergehenden Ausleseverfahrens. Im Abschnitt 1.7 konnte gezeigt werden, daß die Validität von Lehrereinschätzungen zwar im allgemeinen geringer ist als die von Tests (siehe auch Ingenkamp, 1971; Undeutsch, 1969), Grundschulgutachten und Aufnahmeprüfungen sagen den späteren Schulerfolg dennoch besser voraus als eine zufällige Auslese.

Zum anderen gibt es schulische Entscheidungen, wo die Bewerber um einen Platz keine Zufallsauslese ihres Jahrgangs darstellen. Die Validität des Tests darf in solchen Fällen also nicht an dem betref-

fenden Schülerjahrgang gemessen werden, sondern an der Bewerbergruppe.

Beispielsweise bewerben sich an einer Hochschule für Technik sicherlich vor allem Leute, die eine entsprechend ausgerichtete Fachoberschule besucht haben und häufig den Gesellenbrief eines technischen Berufes besitzen. Die Bewerbergruppe ist also ohne irgendeine Auslese schon geeigneter für ein Fachhochschulstudium als der Durchschnitt der gesamten Schülergruppe.

Das Testen von Schülern ist schließlich eine Kostenfrage (Material, Zeit der Schüler, qualifizierte Testleiter, Testauswertung, Beratung). Gleiche ich die geringe prognostische Gültigkeit eines Tests durch eine höhere Selektionsquote aus, bedeutet das: es müssen mehr Schüler getestet werden. Die Verbesserung der Qualität des Testverfahrens ist also (auch) aus diesem Grunde vorzuziehen.

In dem Modell von Cronbach und Gleser ist der Nutzen der Testanwendung dann am höchsten, wenn die Selektionsquote 50% beträgt und der Test eine hohe Validität hat. (In die Definition des »Nutzens« sind die Kosten bereits eingegangen.)

Als Vorgehen mit hohem Nutzen kann man die »sequentielle Strategie« ansehen: aufgrund der Ergebnisse teilt man die Bewerber in drei Gruppen ein: »Angenommen«, »Abgelehnt« und »Weiter untersuchen«. Die letztgenannte Gruppe wird einem weiteren Test unterzogen, diese Schüler werden dann mittels einer Kombination beider Tests beurteilt (Abbildung 6).

Tröger und Lüking (1977) lehnen das Kosten-Nutzen-Modell von Cronbach und Gleser für schulische Fragestellungen ab, weil bei den meisten schulischen Entscheidungen die »Abgelehnten« gleichzeitig in einem anderen Bildungsgang »angenommen« werden müssen. Im Gegensatz zu Einstellungsverfahren in der Wirtschaft kann man die »Abgelehnten« also nicht vergessen. Außerdem wenden Tröger und Lüking ein, daß die Schule weniger den institutionellen als vielmehr den individuellen Nutzen (den Nutzen des Schülers) berücksichtigen sollte.

Schmid (1976) versucht, einen Zusammenhang zwischen der wirtschaftlichen Entwicklung der Bundesrepublik und der Verbreitung von standardisierten Testverfahren (sowie der objektiven Leistungsmessung durch informelle Tests) nachzuweisen. Er stellt fest, daß frühe Ansätze einer deutschen Testpsychologie nach dem zweiten Weltkrieg zunächst unbeachtet blieben und die »verstehende Psychologie« in den 50er Jahren vorherrschte. Diese Zeit war auch die Blütezeit der projektiven Testverfahren (Tests wie der Rorschach-Test, der Wartegg-Zeichen-Test, die metrischen Anforderungen nicht genügten und auch nicht den Anspruch stellten;

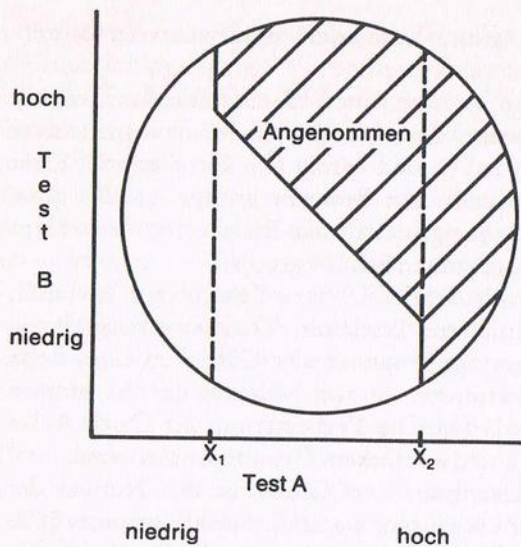


Abbildung 6: Zweistufige sequentielle Strategie.

X_1 : kritischer Testwert – unterhalb dieses Wertes erfolgt eine sofortige Ablehnung der Bewerber.

X_2 : kritischer Testwert – oberhalb dieses Wertes erfolgt eine sofortige Annahme der Bewerber.

Über die Annahme von Bewerbern zwischen X_1 und X_2 wird aufgrund der Kombination der Testwerte von A und B entschieden.

Verfahren, bei denen Reaktionen von Testprobanden gedeutet werden, siehe Heiss (Hrsg., 1963, Kap. 15–18). Zu Beginn der 60er Jahre, so Schmid, wurden qualifizierte Arbeitskräfte knapp, die Folge war eine Unzufriedenheit mit dem Schulsystem und seinen unzulänglichen Auslesemechanismen. Parallel dazu setzte sich in der Psychologie der exakt nach testtheoretischen Gesichtspunkten konstruierte Test durch. Er sollte helfen, »Begabungsreserven« zu entdecken.

Betrachtet man den Einsatz von standardisierten Tests in der Schule unter diesem Gesichtspunkt, versteht man ihn als Mittel zur Vermehrung des Nutzens für die Institution (Schule).

Hinzufügen kann man noch, daß die einsetzende Bildungswerbung zur Überwindung des »Bildungsnotstandes« (siehe Picht, 1965) dazu führte, daß die Prozentanteile der Realschüler und Gymnasiasten an der gesamten Schülerzahl erheblich stiegen. Nun können wir an unsere entscheidungstheoretischen Überlegungen anknüpfen (Abschnitt 1.4 dieses Kapitels) und feststellen: erhöht man die Selektionsquote bei gleicher Gültigkeit des Auslesekriteriums, steigt der Anteil der zugelassenen Ungeeigneten im Verhältnis zu

den akzeptierten Geeigneten an. Das bedeutet: es kommen mehr später erfolgreiche Schüler in die »höheren Schulen«, aber auch sehr viele, die später scheitern werden (siehe Abbildung 2 und 3). Die Öffnung der Realschulen und Gymnasien für mehr Schüler war zur Ausschöpfung von Begabungsreserven also nur von begrenztem Wert.

Die vergleichsweise geringe Gültigkeit von Ausleseentscheidungen mag mit der Benachteiligung der sozialen Grundschicht zusammenhängen, die Tent (1969) nachweisen konnte. Tent stellte aufgrund einer umfangreichen Untersuchung mit dem »Leistungsprüfsystem« (LPS) von Horn fest, daß dieser Test wesentlich schichtunabhängiger ist. Das LPS enthält Aufgabenreihen, die keine sprachlichen Anforderungen stellen. Kemmler (1967) fand dagegen, daß das Lehrerurteil über die Leistungsfähigkeit von Schülern vor allem sprachliche Leistungen und hier besonders die Rechtschreibung berücksichtigt. Geht man von der Annahme aus, daß verbale Leistungen stärker durch den häuslichen Hintergrund geprägt sind, kann die größere Schichtunabhängigkeit des LPS nicht verwundern.

Auch von daher erklärt sich die zunehmende Attraktivität von Tests als Entscheidungshilfe.

(An dieser Stelle muß angemerkt werden, daß Gruppenintelligenztests wie das LPS oder das PSB keineswegs von der Schichtzugehörigkeit unabhängige Testwerte liefern, wie zum Beispiel Kastner, 1978, zeigen konnte.)

Bei pädagogischen Entscheidungen ist der Nutzen des Einsatzes von Tests abhängig von der Reversibilität der einmal eingeschlagenen Schullaufbahn.

Wenn der Schüler nach einer einmal getroffenen Entscheidung für einen Bildungsgang zu jedem Zeitpunkt wechseln kann, ohne Einbußen an Zeit oder zusätzliche Belastungen zu erleiden, ist der Einsatz von Tests ohne Nutzen, denn eine Fehlentscheidung hat keine negativen Folgen für den Schüler. Hat die einmal getroffene Entscheidung allerdings zur Folge, daß andere Bildungsgänge und -abschlüsse dadurch ausgeschlossen werden, sind Fehlentscheidungen für an und für sich geeignete Schüler (Abbildung 2 und 3, Gruppe D) von großer Tragweite. Hier ist der valide Test für den Schüler von großem Nutzen, ebenso wie für die Institution, die »Begabungsreserven« konsequent erreichen will.

Eine irreversible (nicht umkehrbare) Schullaufbahnentscheidung verursacht der Institution Schaden, wenn viele Ungeeignete in einen Bildungsgang geraten (Abbildung 2 und 3, Gruppe B). Sie wird daher den validen Test begrüßen. Ist die Entscheidung dage-

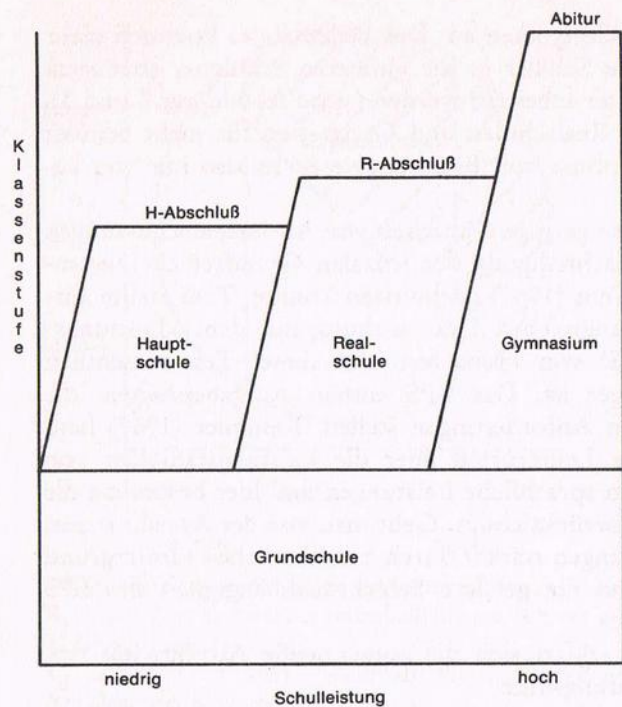


Abbildung 7: Schematische Darstellung eines vertikalen Schulsystems: Für den Schüler ist die Schullaufbahnentscheidung irreversibel – er kann nicht aufsteigen. Für die Institution ist sie dagegen reversibel, sie kann irrtümlich aufgenommene ungeeignete Schüler absteigen lassen.

gen reversibel, kann der Schüler den Bildungsgang jederzeit verlassen und einen geeigneteren besuchen, ist der Test ohne Nutzen.

Das vertikal gegliederte Schulsystem ist grundsätzlich so konstruiert, daß die Übergangsentscheidung von der Grundschule zu den weiterführenden Schulen Hauptschule, Realschule, Gymnasium relativ irreversibel ist. Zumindest gilt das in der einen Richtung: es ist ein Ausnahmefall, wenn ein Schüler ohne Zeitverlust nach dem Übergang am Ende der Grundschul-(oder Orientierungsstufen-) Zeit von der Hauptschule in die Realschule oder von der Realschule in das Gymnasium gelangt. Selbst mit der Wiederholung eines Schuljahres gelingt nur sehr wenigen die Überführung in eine Schulart mit längerem Bildungsgang. In der für den Schüler attraktiven Richtung ist das vertikale System also eher irreversibel.

Die Institution Schule befreit sich von den aufgenommenen Ungeeigneten dadurch, daß sie scheiternde Schüler nicht versetzt und bei einem erneuten Verfehlen des Klassenzieles den weiteren Besuch dieser Schulart verwehrt. Eine gewisse Reversibilität »nach unten«,

also in Richtung auf Schularten mit zeitlich kürzerem Bildungsgang, ist also gegeben (Abbildung 7).

Die Selektionsquote der einzelnen Schularten ist beim Eingang also relativ hoch – durch weitere Entscheidungen im Verlaufe der Zeit wird sie immer niedriger. Ein Teil der Gymnasiasten erreicht das Abitur nicht (verläßt die gymnasiale Oberstufe, verläßt das Gymnasium mit der Realschulreife oder wechselt auf die Realschule über), ein Teil der Hauptschüler verläßt die Schule ohne Abschluß. Ein valider Test würde diese fortlaufenden Selektionsentscheidungen weitgehend überflüssig machen – da keine Ungeeigneten in großer Zahl mehr aufgenommen und keine Geeigneten abgelehnt werden, erreichen mehr Schüler den erstrebten Abschluß (das gilt allerdings nur unter der Voraussetzung, daß die einzelnen Schularten ein festes Leistungsniveau haben und die Auslese nicht deswegen betreiben, weil die Zahl der Plätze in den Abschlußklassen eine unveränderliche Größe ist. Dann nämlich würde der Eingangstest zu einer Heraufsetzung der Selektionsquoten führen).

Im horizontalen System gibt es im Prinzip keine irreversiblen Schullaufbahnentscheidungen, sofern die drei herkömmlichen Schularten nicht durch entsprechende vertikale Leistungsgruppierungen reproduziert werden (»Streaming«, siehe Morrison und McIntyre, 1969, 1971). Horizontal ist die integrierte Gesamtschule. Hier wird die Entscheidung über schulische Abschlüsse erst kurz vor Beendigung der Schulpflicht getroffen. Da zum Zeitpunkt des Eintritts in die Gesamtschule keine Entscheidungen gefällt werden müssen, ist auch ein Test mit hoher Gültigkeit ohne Nutzen. Ob nicht auch im horizontalen System ähnlich verfahren wird, wie für das vertikale System dargestellt, muß man allerdings für die einzelne Schule überprüfen: wenn etwa nach dem »Setting«-System differenziert wird (Leistungskurse für verschiedene Fächer), bedeutet ein Abstieg in Mathematik in einen Kurs mit geringerem Anspruch oft eine endgültige Entscheidung, weil die Differenz im Lernangebot zwischen beiden Kursen nicht mehr aufholbar ist.

2. Das Kriterium Leistung

2.1 Zum Begriff »Leistung«

Hobel (1975) weist aufgrund einer lexikalischen Analyse des Begriffs »Leistung« nach, daß in den pädagogischen Leistungsbegriff verschiedene Aspekte eingegangen sind:

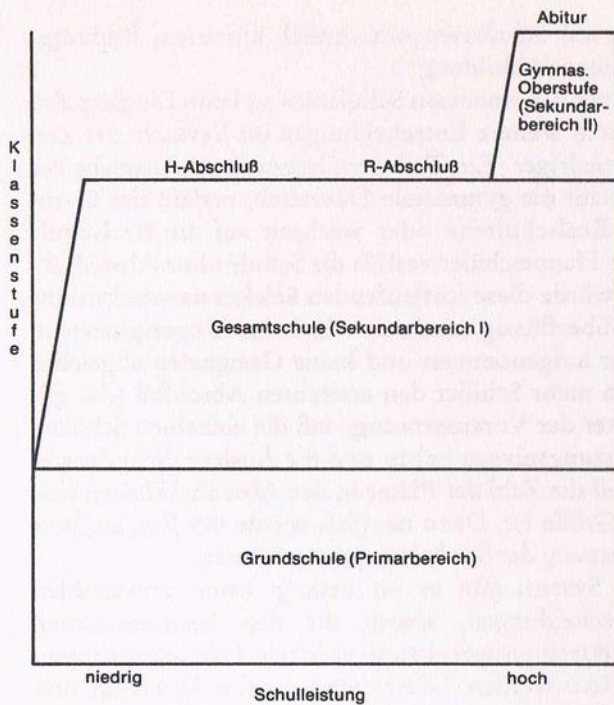


Abbildung 8: Schematische Darstellung eines horizontalen Schulsystems. Schullaufbahnentscheidungen brauchen bis kurz vor dem Ende des Sekundarbereichs I nicht getroffen zu werden. H = Hauptschule; R = Realschule.

- das Moment des Erfüllens einer Pflicht,
- das Moment des Tätigseins, der Tat, wobei Leistung als Prozeß, aber auch als Ergebnis eines Prozesses verstanden werden kann,
- das Moment der Übernahme einer Verpflichtung (Leistungsbereitschaft),
- das Moment der Bewährung einer Fähigkeit (dadurch verbindet sich Leistung mit einer positiven Bewertung,
- das Moment der Zielorientierung.

Hobel vermutet, daß diese verschiedenen Aspekte oft in unterschiedlicher Gewichtung in den Leistungsbegriff eingehen, so daß Mißverständnisse entstehen können. Daher fordert sie, daß der Begriff Leistung nur im Zusammenhang mit einer eindeutigen Definition gebraucht werden sollte. Das gilt insbesondere deswegen, weil in Lehrpläne immer häufiger Leistungsaussagen eingehen.

Krapp und Mandl (1976) unterscheiden zwischen der Schulleistung als hypothetisches Konstrukt im Sinne einer Fähigkeit und dem Schulleistungsverhalten, das direkt beobachtbar ist.

2.2 Leistung als Fähigkeit oder Fertigkeit

Schulleistung kann man deskriptiv als Schülermerkmal definieren, wobei das Zustandekommen der Leistung unberücksichtigt gelassen wird. Sichtbar wird die Schulleistung durch Meßwerte von Tests, im allgemeinen aber durch Lehrerbeurteilungen in Form von Schulnoten (Zensuren). Strenggenommen wird also nicht die Leistung qualitativ als Lernergebnis beschrieben, sondern in bezug auf einen Gütemaßstab bewertet. Dieser Maßstab kann eine vorher festgelegte Lernanforderung sein, aber auch ein Bezugspunkt innerhalb der Lerngruppe. So stellte Ingenkamp (1971) beim Vergleich eines standardisierten Rechen- und eines Englischtests fest, daß Testleistung und Schulnote innerhalb der einzelnen Klassen hoch korrelierten (die Schülerrangreihen nach Leistungstest und Schulnote waren sehr ähnlich). Verglich man aber die Schulnoten der verschiedenen Klassen (zum Beispiel alle »Zweien«) mit den Testergebnissen, war fast kein Zusammenhang zwischen Testergebnissen und Zensuren festzustellen. Eine »Vier« in der einen Klasse konnte mit derselben Testleistung korrespondieren wie eine »Zwei« in einer anderen.

Schulzensuren korrelieren im allgemeinen recht hoch miteinander, wie Tent (1969) für das 6. und Bierhoff-Alfermann (1976) für das 9. und 10. Schuljahr der verschiedenen Schularten feststellen konnten. Beziehungen zwischen einzelnen Zensuren und Testwerten sind dagegen selten. Man könnte sich also auf den Standpunkt stellen, daß Schulnoten vor allem eine globale Fertigkeit widerspiegeln. (Einwände gegen diese Betrachtungsweise sind allerdings:

- Manchmal korrelieren Daten nicht, weil sie inhaltlich etwas Gemeinsames erfassen, sondern weil sie auf die gleiche Weise, mit derselben Methode, gewonnen wurden.
- In die Zensuren gehen Beurteilungsfehler der Lehrer ein wie der »Halo-Effekt« (siehe Hasemann, 1964, S. 827), nämlich von einer Zensur auf die anderen zu schließen. Bei Tent (1969) korrelieren die Schulnoten hoch mit der Beurteilung in »Fleiß«, »Aufmerksamkeit« und »Ordnung«, bei Hadley (1971) mit der Beliebtheit der Schüler bei den Lehrern. Dadurch werden tatsächliche Leistungsunterschiede eines Schülers in verschiedenen Fächern verwischt).

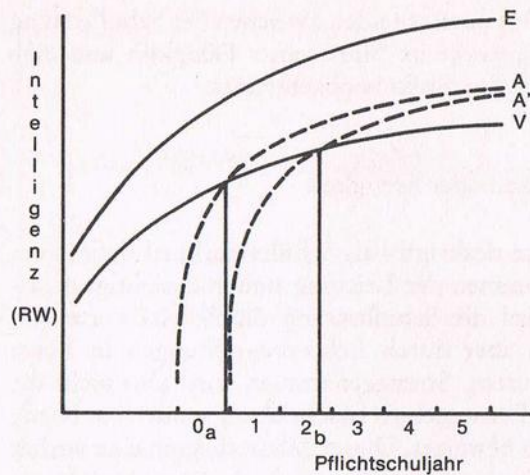


Abbildung 9: Kognitives Grundschulversagen (nach Tiedemann, 1975).

E = erfolgreicher Grundschüler

V = Schulversager

A und A': Anforderungsschwelle

a = 1. Versagen

b = 2. Versagen

RW = Rohwert eines Intelligenztests (absolut, also nicht auf die Altersgruppe bezogen)

Von der Vorstellung eines allgemeinen Schulleistungsniveaus geht das Modell des »kognitiven Grundschulversagens« von Tiedemann (1975) aus. Tiedemann trug verschiedene Ergebnisse über das Versagen in der Schule zusammen und stellte fest, daß ein Sitzenbleiben um so früher erfolgt, je schwächer das Intelligenzniveau eines Schülers ist. Auch die Schwere des Schulversagens (schwache Leistungen, Sitzenbleiben, zweimaliges Sitzenbleiben) korrespondiert mit Intelligenzschwäche. Kemmler (1967) hatte zudem festgestellt, daß Kinder, die bei der Schulreifeuntersuchung zurückgestellt wurden, später meist wieder versagten.

Tiedemann erklärt diese Ergebnisse mit zwei Größen, der Intelligenz des Schülers und einer klassenstufenabhängigen Anforderungsschwelle (Abbildung 9).

Die Intelligenzleistung (Rohwert in einem Test) ist eine negativ beschleunigte Funktion des Lebensalters. Die Intelligenz des Schülers E (Abbildung 9) ist nicht nur höher, sie entwickelt sich auch steiler und erreicht später ein Plateau als die Intelligenz des Schülers V.

Der Schüler V scheitert bei der Einschulungsuntersuchung (a), er

unterschreitet die Mindestanforderung. Die Zurückstellung bedeutet für ihn eine Verschiebung der Anforderungsschwelle um ein Jahr (A'). Da die Kurve der Mindestanforderung aber steiler ist als die Intelligenzentwicklungskurve von V, kommt es zum Zeitpunkt b, am Ende des 2. Pflichtschuljahres, also der ersten Klasse, erneut zum Scheitern.

Tiedemann will mit diesem Modell aufzeigen, daß Zurückstellungen und Klassenwiederholungen relativ unwirksame Maßnahmen sind, um auf Dauer ein Lernversagen zu verhindern.

Tiedemann selbst und in einer Replik Krapp (1977) zeigen auch die Beschränkungen dieses Modells:

- es vernachlässigt, daß leistungsschwache und leistungsstarke Grundschüler überlappende Intelligenzverteilungen aufweisen (Kemmler, 1967);
- einseitige Begabungen sind nicht berücksichtigt;
- ein fest definiertes Anforderungsniveau gibt es nicht, es wechselt vielmehr von Lehrer zu Lehrer, von Schule zu Schule. Das Modell negiert auch die Möglichkeit, durch gezielte pädagogische Behandlung das Leistungsvermögen des Schülers zu verbessern.
- Krapp möchte den Begriff »Intelligenz« oder »kognitive Leistungsfähigkeit« durch »kumulatives Schulleistungsverhalten« ersetzen. Er drückt damit aus, daß der Leistungsstand in der Schule eine Funktion der Lerngelegenheiten für den Schüler ist.

Mit dem letzten Vorschlag drückt Krapp sein Unbehagen an der nur deskriptiven Betrachtungsweise der Schulleistung aus.

2.3 Schulleistung als Verhalten und ihr Zustandekommen

Harnischfeger und Wiley (1977) stellen im Anschluß an Carroll und Bloom fest, daß die aufgewendete Zeit für eine Aufgabe eine entscheidende Determinante der Schülerleistung ist. Der Unterschied in der Lerngeschwindigkeit zwischen Schülern läßt sich einmal dadurch erklären, daß Schüler hinsichtlich des Anteils der aktiven Lernzeit an der insgesamt zur Verfügung stehenden Zeit variieren; zum anderen dadurch, daß die Schüler ihre aktive Lernzeit unterschiedlich gut ausnutzen (Lernrate). Sowohl die Lernrate als auch die aktive Lernzeit ist von Lehrer-(und Unterrichts-) und Schülercharakteristika abhängig (Abbildung 10).

Das Modell von Harnischfeger und Wiley macht deutlich, daß Begabung oder Intelligenz, aber auch früher erworbene Leistungen allein zur Erklärung eines vorliegenden Leistungsergebnisses nicht ausreichen. Entsprechend muß auch die Vorhersagegültigkeit von

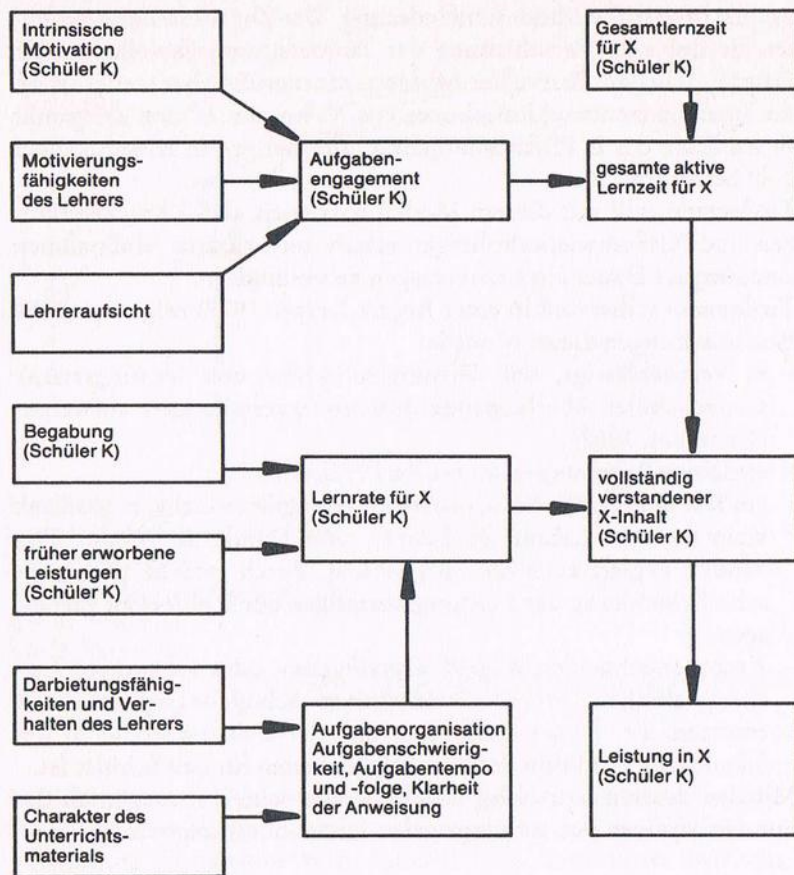


Abbildung 10: Lernfaktoren (aus Harnischfeger und Wiley, 1977) X = Schulfach.

Fähigkeitstests (und Leistungstests) begrenzt bleiben. So ist auch Heller (1978) der Ansicht, daß eine weitere Lernerfolgsprognose die konkreten Lernbedingungen in Gegenwart und Zukunft des Schülers mit umfassen muß – eine gigantische Aufgabenstellung, wenn man sich etwa einen Vergleich zwischen den unterrichtlichen und sozialpsychologischen Gegebenheiten der abgebenden und aufnehmenden Schule in Beziehung zu Schülermerkmalen vorstellt.

2.4 Soziale Herkunft als Determinante des Schulerfolgs

Als Hintergrundbedingung für Schülermerkmale wie »Begabung«, »Motivation«, »früher erworbene Leistungen« ist die soziale Herkunft anzusehen. Müller und Mayer (1976) verglichen die Berufe

von Vätern und Söhnen der Jahrgänge 1920–1936 und kamen zu dem Ergebnis, daß schulische und außerschulische Einflüsse bei der Berufszuweisung die gleiche Rolle spielen. Die Ergebnisse der Pfadanalyse (Müller, 1976) zeigten außerdem, daß innerhalb der »white-collar«-Berufsgruppen eine hohe Mobilität herrscht, daß aber die Schranke von Arbeitern zu Nicht-Arbeitern kaum übersprungen wird (Ausnahme ist die Arbeiterelite). Diese Barriere kommt weitgehend durch außerschulische Einflüsse zustande. Die Institution Schule ist offenbar ungeeignet, dem entgegenzuwirken.

3. Testverfahren für schulische Fragestellungen (eine Auswahl)

Im folgenden sollen einige Testverfahren dargestellt werden, die häufig in Schulpsychologischen Diensten Verwendung finden. Es handelt sich um eine kleine Auswahl aus den vielen deutschsprachigen Tests, die zur Zeit erhältlich sind. Eine ausführlichere Testdarstellung findet sich bei Weise (1975).

3.1 Begabungs-, Fähigkeitstests

3.1.1 Prüfsystem für Schul- und Bildungsberatung von Horn (1969)

Das Prüfsystem für Schul- und Bildungsberatung (PSB) ist ein mehrdimensionales Testverfahren zur Feststellung der Begabungshöhe und -struktur. Ihm liegt das Faktorenmodell von Thurstone (siehe Abschnitt 1.1) zugrunde – gleichzeitig stellt Horn aber auch als Vorteil des Tests heraus, daß die einzelnen Untertests einen hohen Zusammenhang mit dem Gesamtwert zeigen, was ja eher für ein g-Faktoren-Modell spricht.

Das PSB ist eine Kurzform des LPS (Leistungs-Prüfsystem) vom gleichen Verfasser. Normiert wurde das PSB in Baden-Württemberg. Die Tests des LPS wurden ausgewählt, die anhand einer Diskriminanzanalyse am besten zwischen guten und schwachen Schülern unterscheiden.

Die Untertests 1 + 2 sollen den Wortschatz erfassen – für jüngere Schüler geht auch die Rechtschreibleistung mit ein:

Beispielaufgabe: Kräide

In jedem Wort soll der eine falsche Buchstabe durchgestrichen werden.

Die Untertests 3 + 4 repräsentieren den Faktor »Reasoning«:

Beispielaufgabe Test 3: 0 0 0 0 0 0 0 0

Test 4: a b a b a b a á a

5. Zusammenfassung

Intelligenztests unterscheidet man nach der ihnen zugrundeliegenden Faktorentheorie – bei der Zweifaktoren-Theorie geht man davon aus, daß verschiedene Testaufgabentypen einen allgemeinen Begabungsfaktor und daneben spezielle Faktoren erfassen. Die Multiple-Faktoren-Theorie besagt, daß es mehrere voneinander unabhängige Begabungsdimensionen gibt.

Die Schule ist an Tests interessiert, weil sie sich von ihnen präzisere Aussagen erhofft als schulische Beurteilungen liefern können. Tatsächlich werden Tests sehr sorgfältig konstruiert (Aufgabenanalyse). Ein Testautor muß nachweisen, daß sein Verfahren angemessene Normen hat, daß es objektiv auswertbar ist, zuverlässige Meßwerte liefert und wirklich das mißt, was es zu messen vorgibt (Gültigkeit). Die Gütekriterien des Tests allein sagen aber noch nicht aus, ob der Test für schulische Fragestellungen wirklich brauchbar ist.

Ein Problem ist, daß Testnormen schnell veralten. Man kann aber nicht davon ausgehen, daß die 9jährigen Kinder des Jahres 1969 oder 1954 mit den heutigen 9jährigen hinsichtlich ihrer Fähigkeits- und Persönlichkeitsmerkmale identisch sind.

Ob ein Test mit einer gegebenen Gültigkeit eine schulische Entscheidung wirklich verbessert, hängt davon ab, wieviele Bewerber für einen Bildungsgang aufgenommen werden können. Werden praktisch alle Bewerber akzeptiert, ist auch ein Test mit hoher Gültigkeit wertlos – werden dagegen nur sehr wenige Bewerber angenommen, kann man auch mit einem relativ wenig gültigen Test noch mehr Geeignete auslesen als nach Zufall zu erwarten wären. Allerdings lehnt man unter dieser Bedingung auch viele an und für sich Geeignete ab. Ein solches Vorgehen bringt nur der aufnehmenden Institution Vorteile und ist daher problematisch.

Sind mehrere Testverfahren mit geringer Gültigkeit vorhanden,

kann man die Gesamtgültigkeit durch Kombination der Testwerte erhöhen – je nach Entscheidungssituation wählt man die multiple Schnittpunktmethode: Bewerber für einen Bildungsgang müssen in allen Testverfahren Mindestwerte erreichen, wollen sie aufgenommen werden. Bei der linearen Kompensationsmethode kann man ein schwaches Ergebnis in einem Verfahren durch ein gutes in einem anderen ausgleichen.

Viele schulische Entscheidungen lassen sich eher als Klassifikation (mehrere Alternativen) denn als Selektion (zwei Alternativen) beschreiben. Es geht dann darum, Testwertkombinationen und -gewichtungen zu finden, die am besten zwischen Schülern unterschiedlicher Bildungsgänge unterscheiden.

Den Nutzen des Einsatzes von Tests kann man nicht daran messen, was eine zufällige Zuordnung von Schülern zu Bildungsgängen bewirkt hätte, denn auch die schulischen Bewertungen, die zu Entscheidungen führen (zum Beispiel Aufnahmeprüfungen oder Gutachten der abgebenden Schule), haben eine Vorhersagegültigkeit. Außerdem muß man die Kosten (Zeit, Personal, Sachmittel) für die Testung berücksichtigen. Die Senkung der Kosten kann durch sequentielles Vorgehen erreicht werden.

Die Verbreitung standardisierter Fähigkeitstests im schulischen Raum ging einher mit dem gesellschaftlichen Bedürfnis, »Begabungsreserven« zu mobilisieren. Die unzureichende Gültigkeit bisheriger Auslesepraktiken beruhte auch auf ihrer Benachteiligung der sozialen Grundschicht. Bei einer Auslese durch Tests ist diese Benachteiligung nicht in dem Maße gegeben.

Der Nutzen von Tests ist in einem vertikal gegliederten Schulsystem mit geringer Durchlässigkeit von »unten« nach »oben« besonders hoch – erspart er doch Schülern, die für eine Schulart ungeeignet sind und die bisher aufgenommen wurden, das Scheitern (zum Beispiel in Form der zweimaligen Nichtversetzung). Im vertikalen Schulsystem besteht nämlich die Tendenz, Schüler zunächst recht großzügig in die weiterführenden Schulen aufzunehmen, um durch ständiges Senken der Selektionsquote die Ungeeigneten wieder abzugeben.

Im horizontalen Schulsystem (integrierte Gesamtschule) spielen Fähigkeitstests eine viel geringere Rolle, weil irreversible Schullaufbahnentscheidungen nicht so früh gefällt werden müssen (ob dem in der Realität so ist, hängt allerdings von der Art der Differenzierung ab).

Der Begriff »Leistung«, dessen Inhalt der Test ja erfassen soll, wird oft unscharf definiert. Leistung wird oft in Form der Lehrerbeurteilung (Zensuren) ausgedrückt. Die metrischen Eigenschaften von

Zensuren sind ungünstig, Schulnoten aus verschiedenen Klassen wenig vergleichbar.

Modelle der Leistungsentwicklung orientieren sich oft an der Entwicklung der Intelligenz. Tatsächlich lassen sich mit solchen Modellen manche schulischen Phänomene wie die relative Unwirksamkeit des Sitzenbleibens als Fördermaßnahme erklären. Der Nachteil solcher Modelle ist, daß die Leistungsfähigkeit des Schülers als unwandelbare Größe und nicht beeinflussbar durch pädagogische Bemühungen gesehen wird.

Die Leistung eines Schülers ist eben nicht nur durch seine Begabung und seine bisherigen Lernerfahrungen bestimmt, sondern auch durch Lehrer- und Unterrichtsmerkmale sowie die aktive Lernzeit des Schülers.

Das Prüfsystem für Schul- und Bildungsberatung ist ein mehrdimensionales Testverfahren und weist für schulische Beratungszwecke verschiedene Gültigkeitsaspekte auf. Der CFT 2 soll dagegen nur relativ sprachunabhängige Intelligenzmerkmale erfassen. Ob der CFT 2 wirklich milieuunabhängiger als das PSB ist, scheint aber nicht sicher.

Leistungstests sollen unterrichtsbezogene Inhalte prüfen und Leistung präziser messen als die Schulzensuren. Die Allgemeinen Schulleistungstests (hier AST 2 und 4) erfassen die Leistungen mehrerer Fächer. Die Autoren können allerdings nicht belegen, daß sie wirklich verschiedene Leistungsmerkmale messen. Für viele Einzelaspekte der Leistung liegen spezielle Testverfahren vor. Auch im Persönlichkeitsbereich gibt es nach den Prinzipien der Testkonstruktion erstellte Verfahren, die Angst, Schulangst, Schulunlust, Anstrengungsvermeidung und Pflichteifer erfassen.

Dem Schulpsychologen bleibt es in der individuellen Beratung letztlich überlassen, Persönlichkeitsaspekte in die Leistungsprognose einzubeziehen. Bei der Schullaufbahnberatung kann es auch nicht darum gehen, den Schüler alternativ als »geeignet« oder »nicht geeignet« einzuordnen, sondern ihm eine Wahrscheinlichkeit zu nennen, mit der er zu den Geeigneten zählt und ihm und den Eltern die Entscheidung zu überlassen. Standardisierte Fähigkeitstests (und meist auch Leistungstests) verschaffen dem Berater nur in begrenztem Umfang Veränderungswissen. Im Falle der Lese-Rechtschreibschwäche läßt sich zeigen, daß die Feststellung von Fähigkeitsdefiziten der Legastheniker nicht zwingend zu Therapievor schlägen führt. Dazu wäre es vielmehr notwendig, die Lernprozesse erfolgreicher und wenig erfolgreicher Schüler zu vergleichen.

Literatur

- Angermaier, M. (1974): Psycholinguistischer Entwicklungstest – Manual. Beltz, Weinheim.
- Angermaier, M. (1974): Sprache und Konzentration bei Legasthenie. Hogrefe, Göttingen.
- Bachmair, G. (1976): Methoden und Konsequenzen pädagogischer Entscheidungen. Deutsche Schule, S. 639 – 654.
- Bierhoff-Alfermann, D. (1976): Die Beziehung von Noten und Schülermerkmalen bei Schülern der 9. + 10. Klasse. Psychologie in Erziehung und Unterricht, 23, S. 205 – 214.
- Bondy, C. (Hrsg.), Hardesty, F. P., Priester, H. J. (1956): Hamburg-Wechsler-Intelligenztest für Kinder. Huber, Bern, Stuttgart.
- Brickenkamp, R. (1970³): Test d2, Aufmerksamkeitsbelastungstest. Hogrefe, Göttingen.
- Cohen, J. (1959): The factorial structure of the WISC at ages 7;6, 10;6 and 13;6. Journ. consult. Psychol. 23, S. 285 f.
- Correll, W., Ingenkamp, K. (1958²): Fremdsprachen-Eignungstest für die Unterstufe (FTU 4 – 6). Beltz, Weinheim.
- Cronbach, L. J. (1960): Essentials of psychological testing. Harper, New York.
- Cronbach, L. J., Gleser, G. C. (1965²): Psychological Tests and Personnel Decisions. University of Illinois Press, Urbana.
- Fippinger, F. (1967): Allgemeiner Schulleistungstest für 4. Klassen (AST 4). Beltz, Weinheim.
- Furcht-Krafft, E. (1976): Der Einfluß von Umweltbedingungen auf Intelligenztestergebnisse. Lehren und Lernen, Zeitschr. des Instituts für Bildungsplanung und Studieninformation Stuttgart 8, S. 12 – 21.
- Gebauer, T. (1965): Vergleichende Untersuchung über den Voraussagewert von Aufnahmeprüfung und Testuntersuchung für den Erfolg auf weiterführenden Schulen in: Ingenkamp, K. (Hrsg.) Schulkonflikt und Schülerhilfe. Beltz, Weinheim.
- Hadley, S. T. (1971): Feststellungen und Vorurteile in der Zensurierung. in: Ingenkamp, K. (Hrsg.) Die Fragwürdigkeit der Zensurierung. Beltz, Weinheim.
- Harnischfeger, A., Wiley, D. E. (1977): Kernkonzepte des Schullernens. Zeitschr. f. Entwicklungsps. u. Päd. Ps. IX, 3, S. 207 – 228.
- Hasemann, K. (1964): Verhaltensbeobachtung in: Handbuch der Psychologie Band 6, Psychologische Diagnostik, Hrsg. Heiß, R. Hogrefe Göttingen.
- Heiß, R. (Hrsg.) (1964): Handbuch der Psychologie Band 6, Psychologische Diagnostik. Hogrefe, Göttingen.
- Heller, K. (1976): Computerunterstützte Interpretation von Testbefunden in der Schullaufbahnberatung in: Heller, K. (Hrsg.) Handbuch der Bildungsberatung. Klett, Stuttgart.
- Heller, K. (1978): Ein Prognosemodell für die Schullaufbahnberatung in: Heyse, H. (Hrsg.) Texte zur Schulpsychologie und Bildungsberatung Band 3. Westermann, Braunschweig (in Vorbereitung).
- Hobel, S. (1975): Pädagogischer Leistungsbegriff und Leistungsorientierung in der Schule. Diss., Bielefeld.
- Hofstätter, P. R. (1972): Psychologie, Das Fischer Lexikon, Bd. 6. Fischer Taschenbuchverlag, Frankfurt a.M.
- Horn, W. (1962): Leistungs-Prüf-System (LPS). Hogrefe, Göttingen.

- Horn, W. (1969): Prüf-System für Schul- und Bildungsberatung (P-S-B). Hogrefe, Göttingen.
- Ingenkamp, K. (1971): Sind Zensuren aus verschiedenen Klassen vergleichbar? in: Ingenkamp, K. (Hrsg.) Die Fragwürdigkeit der Zensurengebung. Beltz, Weinheim.
- Jung, U. O. H. (Hrsg.) (1977): Lese-Rechtschreibschwäche und Sprachunterricht. IPTS-Katalog 22, Landesinstitut Schleswig-Holstein für Praxis und Theorie der Schule, Kiel.
- Kastner, M. (1978): Zur Problematik von Tests zum Übergang an weiterführende Schulen. *Psychologie in Erziehung und Unterricht*, 25, S. 9-15.
- Kemmler, L. (1967): Erfolg und Versagen in der Grundschule. Hogrefe, Göttingen.
- Krapp, A. (1977): Schulversagen als Leistungsversagen: Kritik des Tiedemannschen Erklärungsmodells. *Zeitschr. f. Entwicklungsps. u. Päd. Ps.* IX, 1, S. 36-39.
- Krapp, A., Mandl, H. (1976): Vorhersage und Erklärung der Schulleistung. *Zeitschr. f. Entwicklungsps. u. Päd. Ps.* VIII, 3, S. 192-219.
- Krauth, J., Lienert, G. A. (1973): KFA. Die Konfigurationsfrequenzanalyse und ihre Anwendung in Psychologie und Medizin. Alber, Freiburg.
- Küffner, Helmuth, Rothe, Claus (1977): Probleme der Verwendung von Testnormen in der schulpsychologischen Arbeit in: Arnhold, W. (Hrsg.) *Texte zur Schulpsychologie und Bildungsberatung*, Band 2. Westermann, Braunschweig.
- Langfeldt-Nagel, M., Langfeldt, H.-P. (1976): Testverfahren für die Bildungsberatung in: *Handbuch der Bildungsberatung*; Hrsg. Heller, K. Klett, Stuttgart.
- Lawshe, C. H., Bolda, R. L., Auclair, G.: (siehe McCormick und Tiffin, 1974, S. 122; 601 f.).
- Lienert, G. A. (1969³): Testaufbau und Testanalyse. Beltz, Weinheim.
- Magnusson, D. (1975²): Testtheorie. Deuticke, Wien.
- Masendorf, F., Roeder, B. (1974): Typologisierung lernschwacher Schüler mit Hilfe der Konfigurationsfrequenzanalyse (KFA). *Psychologie in Erziehung und Unterricht*, 21, S. 327-334.
- McCormick, E. J., Tiffin, J. (1974⁶): *Industrial Psychology*. Prentice Hall, New Jersey.
- Michel, L. (1964): Allgemeine Grundlagen psychometrischer Tests, in: *Handbuch der Psychologie Band 6, Psychologische Diagnostik*, Hrsg. Heiß, R. Hogrefe, Göttingen.
- Möhling, R., Raatz, U. (1974): Konzentrationstest für 1. Klassen (KT 1). Beltz, Weinheim.
- Mönikes, D., Mönikes, W. (1975): Der HAWIK, ein unzulängliches Prüfinstrument in der Schule für Lernbehinderte im Vergleich zum CFT 2 - eine Voruntersuchung. *Zeitschrift für Heilpädagogik*, 26, 8, S. 435-444.
- Morrison, A., McIntyre, D. (1969): *Teachers and teaching*. Penguin Books, Harmondsworth, Middlesex.
- Morrison, A., McIntyre, D. (1971): *Schools and socialization*. Penguin Books, Harmondsworth, Middlesex.
- Müller, W. (1976): Bildung und Mobilitätsprozeß - Eine Anwendung der Pfadanalyse. *Zeitschr. f. Soziologie*, Jg. 1, 1, S. 65-84.
- Müller, W., Mayer, K. U. (1976): Chancen durch Bildung? Deutscher Bildungsrat - Gutachten und Studien der Bildungskommission - 42 -. Klett, Stuttgart.
- Nickel, H., Schlüter, P., Fenner, H.-J. (1973): Angstwerte, Intelligenztest- und Schulleistungen sowie der Einfluß der Lehrerpersönlichkeit bei Schülern verschiedener Schularten. *Psychologie in Erziehung und Unterricht*, 20, S. 1-13.
- Piaget, J., Inhelder, B. (1972): *Die Psychologie des Kindes*. Walter, Freiburg.
- Picht, G. (1965): *Die deutsche Bildungskatastrophe*. Walter, Olten.
- Priester, H. J. (1964): Intelligenztests für Erwachsene in: *Handbuch der Psychologie Band 6, Psychologische Diagnostik*, Hrsg. Heiß, R. Hogrefe, Göttingen.
- Raatz, U., Möhling, R. (1971): *Frankfurter Tests für Fünfjährige - Konzentration (FTF-K)*. Beltz, Weinheim.
- Rieder, O. (1971): *Allgemeiner Schulleistungstest für 2. Klassen (AST 2)*. Beltz, Weinheim.
- Rollett, B., Bartram, M. (1977): *Anstrengungsvermeidungstest*. Westermann, Braunschweig.
- Royle, W., Schwarzer, R.: Zur säkularen Akzeleration der PSB-Intelligenz. *Diagnostika XXII*, 3, Seite 99-110.
- Sapolski, A.: (siehe Tausch und Tausch, 1970³, S. 79).
- Scheerer-Neumann, G. (1977): Funktionsanalyse des Lesens, Grundlage für ein spezifisches Lesetraining. *Psychologie in Erziehung und Unterricht*, 24, S. 125-135.
- Schmid, R. (1976): Zur Geschichte und Funktion von Leistungsdifferenzierenden Testverfahren unter besonderer Berücksichtigung des schulischen Bereichs. Diss. Bremen.
- Schnorr, I. H. (1976): Zur Frage der Überweisung von Ausländerkindern in die Schule für Lernbehinderte. *Zeitschrift für Heilpädagogik*, 27, 2, S. 79-88.
- Tausch, R., Tausch, A. M. (1970³): *Erziehungspsychologie*. Hogrefe, Göttingen.
- Tent, L. (1969): Die Auslese von Schülern für weiterführende Schulen. Dr. H. C. Hogrefe, Göttingen.
- Tiedemann, J. (1975): Kognitives Grundschulversagen. Eine Denkvorstellung. *Zeitschr. f. Entwicklungsps. u. Päd. Ps.* VII, 4, S. 289-298.
- Tröger, H., Lüking, J. (1977): Validität und Fehlklassifikation bei Selektionsentscheidungen in: Arnhold, W. (Hrsg.) *Texte zur Schulpsychologie und Bildungsberatung*, Band 2. Westermann, Braunschweig.
- Ulmann, G. (1968): *Kreativität. Neue amerikanische Ansätze zur Erweiterung des Intelligenzkonzeptes*. Beltz, Weinheim.
- Undeutsch, U. (1969): Zum Problem der begabungsgerechten Auslese beim Eintritt in die höhere Schule und während der Schulzeit, in: Roth, H. (Hrsg.) *Begabung und Lernen*. Deutscher Bildungsrat. Gutachten und Studien der Bildungskommission, Band 4. Klett, Stuttgart.
- Valtin, R. (1970): *Legasthenie - Theorien und Untersuchungen*. Beltz, Weinheim.
- Wechsler, D. (1956): *Die Messung der Intelligenz Erwachsener* (Hrsg. Bondy, C. Amerikan. Original erschienen 1944). Huber, Bern.
- Weiß, R. H. (R. B. Cattell) (1972): *Grundintelligenztest CFT 2*. Westermann, Braunschweig.
- Wiczercowski, W., Bastine, R., Fittkau, B., Nickel, H., Tausch, R., Tewes, U. (1969): Verminderung von Angst und Neurotizismus bei Schülern durch positive Bekräftigungen im Schulunterricht. *Zeitschr. f. Entwicklungsps. u. Päd. Ps.* 1, S. 3-12.
- Wiczercowski, W., Nickel, H. (1974): *Anstfragebogen für Schüler (AFS)*. Westermann, Braunschweig 1974.