Zeitschrift: Schweizer Schule
Herausgeber: Christlicher Lehrer- und Erzieherverein der Schweiz
Band: 54 (1967)
Heft: 5

Artikel: Psychologische Tests in der Schule
Autor: Flammer, A.
DOI: https://doi.org/10.5169/seals-528477

Psychologische Tests in der Schule I

Dipl. Psychologe A. Flammer, Fribourg

In weiten Kreisen des privaten und öffentlichen Lebens hat der Test längst breiten Eingang gefunden. Skeptiker sprechen von ‹Testitis› und brandmarken damit das blinde Vertrauen, das oft in dieses psychologische Hilfsmittel gesetzt wird, und den Schwund an Einfühlungsgabe und Mut zur persönlichen Begegnung. Sie sehen den getesteten Menschen gestempelt und der Chance beraubt, durch eigene Anstrengung Erwartungen zu übertreffen.

Seit Jahren geht nun auch die Rede von Schultests (oder Schulleistungstest). In den USA wird jeder Schüler etliche Male während seiner Schullaufbahn getestet; und für Deutschland nennt *Ingenkamp* (1962) 45 Schulleistungstests.

Was kann ein Test der Schule bieten? Und ist es sinnvoll, seine Dienste zu beanspruchen? Bevor wir versuchen, Ansätze von Antworten auf diese Fragen zu geben, müssen wir verschiedene

Arten von Tests

unterscheiden. Die wissenschaftliche Literatur nennt eine ganze Reihe von Einteilungen; hier seien drei ausgewählt und kurz dargestellt:

Irle (1956) unterscheidet:

1. *Allgemeine Intelligenztests.* Diese sollen das allgemeine und umfassende Begabungspotential, die ‹Intelligenz›[1], messen.
2. *Fähigkeits- und Begabungstests* messen gesonderte Fähigkeiten (zum Beispiel Handgeschicklichkeit, Hörschärfe) und Begabungen (zum Beispiel Eignung für bestimmte Berufe).
3. *Kenntnis- und Leistungstests* messen ‹Gelerntes› im engeren Sinn.
4. *Persönlichkeitstests* untersuchen Interessen, Neigungen, Konflikte usw.

Lienert (1961, S. 19) nennt drei Gruppen:

1. *Intelligenztests.* Darunter fällt auch die zweite Gruppe von *Irle.*
2. *Leistungstests.*
3. *Persönlichkeitstests.*

[1] *Montalta* (1959, S. 195) definiert Intelligenz als die ›zentrale Fähigkeit, neue Situationen ihrem Wesen gemäß zu bewältigen‹.

Das Handbuch der Psychologie, Band 6 (‹Psychologische Diagnostik›, 1964), faßt auch die ersten beiden Gruppen *Lienerts* zusammen und unterscheidet:

1. *Fähigkeitstests,*
2. *Persönlichkeitstests,*
3. *Sonstige diagnostische Verfahren,* unter die unter andern die Diagnostik der sozialen Beziehungen fällt.

Welche dieser Testarten kommen nun

für die Schule

in Betracht?

Wenn wir den ausgebildeten Schulpsychologen mit einbeziehen, dann wohl alle. Seine Abklärungen (Hilfsschulbedürftigkeit, Betragensschwierigkeiten, Lese-Rechtschreibschwäche usw.) bedürfen mit wechselnder Akzentsetzung des ganzen diagnostischen Instrumentariums.

Hingegen wäre es der Sache wenig dienlich, wollte man auch dem praktischen Lehrer Tests aller Arten zur Anwendung empfehlen. Die Interpretation von *Intelligenztests* bedarf einer gründlichen psychologischen Allgemeinbildung und Erfahrung, und erst recht setzen die meist projektiven[2] *Persönlichkeitstests,* die in Europa üblich sind, eine solide tiefenpsychologische Ausbildung voraus. Leistungstests aber, und zwar *Schulleistungstests*[3], sollen unseres Erachtens unter bestimmten Umständen durchaus in der Hand des Klassenlehrers zur Anwendung kommen. Darüber hinaus möchten wir ihm *soziometrische Methoden* (vgl. *Cappel* 1963; *Heller* 1964 und andere), die unter die dritte Gruppe der Handbuch-Einteilung fallen, sehr empfehlen. Nicht unbedingt ausschließen von der Liste möchten wir *Schulreifetests* (vgl. *Kern* 1958; *Strebel* 1957; FST und andere), die als ‹Entwicklungstests› oft zur Gruppe ‹Intelligenztests› geschlagen werden.

Das wichtigste Kriterium für die Anwendung dieser Methoden ist sicher ein Minimum an Wis-

[2] Projektive Tests sind solche, die den Probanden veranlassen, sich oder Eigenarten seiner selbst in seine Äußerungen zu legen, zu ‹projizieren›. Der bekannteste dieser Tests ist der Formdeutversuch von *H. Rorschach.*

[3] Unter Schulleistungstests verstehen wir mit *Ingenkamp* (1962, S. 16) solche, die die Ergebnisse ‹schulischen Lernens› untersuchen.

sen und Erfahrung und die nötige Kritikfähigkeit. Zum ersten dieser drei Punkte und indirekt auch zum dritten einen kleinen Beitrag zu leisten, ist das Anliegen dieses Aufsatzes. Besonders die Kenntnis der Regeln des wissenschaftlich saubern

Testaufbaus

scheint uns geeignet, sowohl allzu leichtfertiger Testanwendung zu steuern, als auch die grundsätzlichen Gegner dieser ‹Nur-Scharlatanerie› – auch ihrer sind noch viele unter unsern Kollegen – zu einer differenzierteren Stellungnahme zu bewegen. Zu diesem Zweck wollen wir uns auf die Leistungstests beschränken. [4]

Nehmen wir an, es soll ein Schulleistungstest aufgebaut werden, der an Sechstkläßlern die Kenntnis der Grammatik prüfen und damit zum Beispiel einen erfolgreichen Lateinunterricht am Gymnasium vorhersagen helfen soll.

Vorfragen

Meistens kann nicht spontan entschieden werden, wie ein solcher Test aussehen muß, damit er das ihm gesteckte Ziel erreicht. Durch Merkmalsanalysen und Besprechungen mit Fachleuten – *rationale Analyse* – entstehen Mutmaßungen oder Hypothesen, über deren Richtigkeit meist erst der empirische (= praktische) *Gültigkeitsnachweis* nach der Aufbauarbeit entscheiden wird. In unserm Beispieltest könnte man sich zum Beispiel für eine bloße Wissensprüfung entscheiden oder der Untersuchung der richtigen praktischen Anwendung der Grammatik in der Muttersprache den Vorzug geben, oder gar beide Möglichkeiten kombinieren.

Die nächsten Entscheidungen betreffen die Wahl des

Aufgabentypus.

Ein freier Aufsatz zum Beispiel böte den Vorteil der Ungezwungenheit, wäre aber sehr schwierig, gerecht beurteilt zu werden. [5] Man könnte dem Schüler einen fehlerhaften Text zur Bezeichnung oder Berichtigung der Fehler vorlegen, oder eine Aufstellung von Regelsätzen, wobei zum Beispiel je Dreiergruppe einer richtig und anzustreichen wäre. Oft wird in solchen Fällen auch ein Lückkentext vorgegeben oder in leichter Abwandlung davon einer, der in Klammern die Grundform des einzusetzenden Wortes enthält. Es lassen sich noch mehr solcher Formen ausdenken; gegen jede lassen sich irgendwelche Einwände finden: Ratemöglichkeit und unglückliche Vorbildwirkung bei den Mehrfach-Wahl- und Umformungs-Aufgaben, Zweideutigkeiten in den Ergänzungsaufgaben, Unzulänglichkeit für manche Teilgebiete bei der Lückenform usw. Dazu kommen Forderungen (bzw. Wünschbarkeiten) nach Zeitökonomie, geringem Material-, bzw. Papieraufwand, einfacher Durchführbarkeit und anderem mehr.

Nach all diesen Vorentscheidungen kann die eigentliche

Aufgabenkonstruktion

beginnen. Auch wenn nach sorgfältiger rationaler Analyse ein tadellos scheinender Aufgabenentwurf vorliegt, muß der Testautor mißtrauisch sein. Praktische

Einzelversuche

mit den Schülern, die die Aufgaben laut lesen und auch ihre Lösungsversuche aussprechen müssen, werden noch manche Unklarheiten in der Aufgabenstellung, ungeläufige Wörter und unerwünschte Ausweg-Lösungsmöglichkeiten aufzeigen. Immer wieder müssen Aufgaben ersetzt, abgeändert und neu vorgelegt werden.

Mit der nunmehr *empirisch voranalysierten* Aufgabenreihe kann jetzt das Kernstück der ganzen Analyse vorgenommen werden, das oft als die

Aufgabenanalyse

schlechthin bezeichnet wird. Sie besteht darin, daß eine Stichprobe von mindestens 400 Individuen den vorläufigen Test durcharbeitet, und daß mit diesen Ergebnissen eine sog. Schwierig-

[4] Auch bei diesen können wir natürlich nicht alle auch noch möglichen und wünschbaren, zum Teil raffinierten Methoden besprechen, die die moderne Teststatistik bisher ausgearbeitet hat. Der interessierte Leser sei auf die Literaturangaben im Anhang verwiesen, vor allem auf *Lienert* (1962), auf den sich auch diese Ausführungen stark stützen.

[5] Untersuchungen zeigen immer wieder, wie weit die Urteile verschiedener Lehrer über gleiche Aufsätze voneinander abweichen. Vergleiche unter andern *Horney*, 1960, S. 160–161; IMK-Jahresbericht 1966; *Ingenkamp* 1964, S. 122; *Kötter* und *Graul* 1965; *Pally* 1955, S. 93–94; *Samstag* und *Baus* 1962, S. 119–120.

keits- und eine Trennschärfeanalyse durchgeführt wird. Die erstere soll sicherstellen, daß die *Schwierigkeitsgrade* der Einzelaufgaben[6] dem Leistungsniveau der später zu testenden Schülerschaft entspricht. Eine zu schwere Aufgabe, die von keinem Sechstkläßler (in unserm Beispiel) richtig gelöst werden kann, trägt zur Differenzierung der Fähigkeitsstufen der Schüler ebenso wenig bei wie eine Aufgabe, die von sämtlichen Schülern richtig gelöst wird. Zudem sollen die Anforderungen auf der ganzen Fähigkeitsskala der teilnehmenden Schüler gleichmäßig verteilt sein, um jedem gerecht zu werden. Es ist allerdings denkbar, einen Test ausschließlich für die Auswahl der besten 10 Prozent der sechsten Klasse für das Gymnasium zu konstrukturieren und dafür nur Aufgaben zu wählen, die von etwa 10 Prozent aller Schüler richtig gelöst werden (das heißt Aufgaben mit einem Schwierigkeitsindex um 10). In den meisten Fällen aber wird man die Zielsetzung weiter fassen, um breitere Anwendungsmöglichkeiten zu schaffen. In diesem Zusammenhang wird auch verständlich, warum in einem Leistungstest von den getesteten Schülern etwa die Hälfte der Aufgaben gar nicht oder falsch gelöst werden *muß*. Dieser testtheoretischen Forderung nach maximaler Auslastung des Tests muß eine psychologische und pädagogische gegenübergestellt werden: Das Erlebnis des gehäuften Versagens lähmt nicht nur die weitere Leistungsfreude und Leistungskraft im Test (was die Testresultate herabdrückt), sondern ist überhaupt jeglichem Lernen – wofür der Schüler ja in der Schule ist – abträglich; deshalb muß der Testkonstrukteur von Anfang an darauf achten, Aufgabentypen zu finden, die dem versagenden Schüler sein Unvermögen nicht zu deutlich vor die Augen halten. Deshalb sind zum Beispiel Aufgaben, zu denen nur die richtige oder gar keine Lösung geschrieben werden kann, unerwünscht. In diesem Punkt liegt auch einer der Gründe, warum später in der Test-Endform die Aufgaben nach aufsteigender Schwierigkeit gereiht werden.

Der zweite Hauptzweck der Aufgabenanalyse ist die Berechnung der Trennschärfe-Indices jeder Aufgabe. Unter *Trennschärfe* versteht man gemeinhin die Tatsache, daß die Schüler, die eine bestimmte Aufgabe richtig lösen, effektiv fähiger sind als die versagenden und deshalb zu Recht besser qualifiziert werden. Das scheint auf den ersten Anhieb eine triviale Aussage zu sein. Die Teststatistik kann aber nachweisen, daß das Faktorenbündel, das die Lösung einer bestimmten Aufgabe bedingt, oft sehr bunt ist und im Einzelfall bedeutende Elemente enthalten kann, die mit dem betreffenden Test gar nicht zu messen beabsichtigt sind. So kann es einfach durch den Dialekt bedingt sein, daß ein Berner Schüler ‹Ihr› als Fürwort der Höflichkeitsform verwendet, während ein in dieser Beziehung gleich fähiges Kind aus St. Gallen eine solche Versuchung gar nicht empfindet. Oder eine etwas schwierigere Wortform gelingt einer Anzahl von Kindern nur, weil in ihrer Klasse oft ein Lied gesungen wird, das diese Form an exponierter Stelle aufweist. Meist handelt es sich aber um viel weniger auffällige ‹Zufallsfaktoren›, zum Beispiel um eine unklare Aufgabenstellung, die zum bloßen Raten verleitet; und Raten ist an sich für schwache Schüler nicht schwieriger als für gute.[7]

Praktisch wird der Trennschärfe-Index gewonnen durch einen Vergleich zwischen einer Gruppe von Schülern, die im Gesamttest ein gutes Resultat erzielte, und einer gleich großen schwachen Gruppe. Eine einfache der vielen möglichen Formen lautet:

$$TI = \frac{R_o - R_u}{N_t}$$

wobei: TI = Trennschärfe-Index

R_o = Anzahl richtiger Lösungen (der betreffenden Aufgabe), die von der im *Gesamttest* obern Gruppe, zum Beispiel der obersten 30 Prozent, gegeben worden sind.

R_u = Richtige Lösungen der untern Gruppe, zum Beispiel der untersten 30 Prozent.

N_t = Anzahl Versuchspersonen in jeder der *gleich großen* Teilgruppen.

Aus dieser einfachen Berechnungsart der Trennschärfe oder Diskriminierungsfähigkeit jeder Auf-

[6] Die selteneren Testformen, in denen nicht Einzelaufgaben isoliert werden können, können in manchen Punkten nicht gleich analysiert werden. Ihre Behandlungsweise ist analog; es kann hier aber nicht darauf eingegangen werden.

[7] Für Tests mit Auswahl-Antwort-Form, bei der natürlich auch bloßes Raten zu einem gewissen Resultat führen kann (bei Aufgaben mit je drei Möglichkeiten theoretisch zu 33,3% der Maximalpunktzahl!), sind Formeln für die ‹Zufallskorrektur› erarbeitet worden (vgl. *Lienert* 1961, S. 79–80).

gabe geht deutlich ein Grundproblem solchen Testaufbaus hervor. Wenn unser gedachter Test zum Beispiel eine breite Skala menschlicher Fähigkeiten überdeckt, in der eine Fähigkeitsgruppe nicht notwendig auch eine andere bedingt, ist es ‹ungerecht›, die Aufgaben nach den Gesamtresultaten (als Kriterien) zu beurteilen. Entweder werden damit ‹Außenseiterfähigkeiten› oder solche, die im Testentwurf wenig zahlreich vertreten sind, zu schlecht qualifiziert, oder auch der ganze Test erhält durchschnittlich eine zu schlechte Trennschärfequalifikation. Deshalb versucht man heute immer mehr, durch Homogenitätsuntersuchungen und besonders mit sogenannten Faktorenanalysen möglichst eindeutige, ‹gesäuberte› oder homogene Einzeltests aufzubauen, die in der Zusammensetzung einer sogenannten Testbatterie dennoch komplexere Abklärungen ermöglichen. Auch in der praktischen Arbeit des Psychologen und in unserm Fall des Lehrers ist es ja durchaus wünschenswert, möglichst exakt zu wissen, was ein Test mißt und was er eben ausklammert. Andererseits drängt sich heute gerade deshalb eine besondere Warnung vor allzusehr generalisierender, vorschneller Interpretation der Resultate auf.
Nach dieser Aufgabenanalyse setzt nun der Testautor die

Endform

zusammen: Anhand der Schwierigkeitsindices wählt er aus seinen Aufgaben diejenigen aus, die zugleich die geforderte Schwierigkeitsverteilung ergeben und einen hohen Trennschärfe-Index aufweisen. Daß er die Aufgaben nach aufsteigender Schwierigkeit ordnet, wurde schon an anderer Stelle erwähnt. Es ist klar, daß nun von den bisherigen Aufgaben eine ganze Reihe wegfällt. Daran muß schon von Anfang an gedacht werden. Im allgemeinen wird deshalb empfohlen, etwa 200 Prozent der Aufgabenzahl der geplanten Endform in die Aufgabenanalyse zu nehmen.

Mit dem Aufbau der Endform ist die Testentwicklung aber noch nicht abgeschlossen. Als nächster wichtiger Schritt ist die

Zuverlässigkeitsprüfung

zu nennen. Unter Zuverlässigkeit (englisch reliability = Reliabilität) eines Tests versteht man die Tatsache, daß er das *exakt mißt*, was er mißt, und zwar unter Ausklammerung der Frage nach dem Was der Messung. Nach *Lienert* (1961, S. 17) besteht die Zuverlässigkeit aus drei Faktoren. Die *Objektivität* oder Unzweideutigkeit der Auswertung haben wir schon einmal genannt. Oft wird eigens dafür eine empirische Abklärung angestellt, indem unabhängig voneinander mehreren Lehrern gleiche Lösungsversuche von getesteten Schülern zur Bewertung gegeben und ihre Angaben dann verglichen werden. Die *Stabilität* des Tests ist der zweite Faktor und bedeutet ‹Meßgenauigkeit›, Unabhängigkeit von äußern Faktoren, psychologische Identität der Aufgabenstellung in jedem Testfall. Der dritte Faktor, eine minimale *Konstanz des Persönlichkeitsmerkmals*, ist überhaupt Voraussetzung zum Nachweis der Zuverlässigkeit.

Die Zuverlässigkeit wird üblicherweise dadurch gemessen, daß der gleiche Test nach einem bestimmten Zeitabstand den gleichen Versuchspersonen nochmals zur Bearbeitung vorgelegt wird (Re-Test-Methode). Die zu vergleichenden Resultate sollen dabei möglichst die gleichen bleiben. Diese Methode ist jedoch nur durchführbar, wenn die Testaufgaben nicht die Tendenz haben, wegen besonderer Aktualität oder Eigenart im Gedächtnis leicht haften zu bleiben. Die Verlängerung des zeitlichen Zwischenraumes kann manchmal abhelfen, vermindert aber die Konstanz des Persönlichkeitsmerkmals. Die Testtheorie bietet aber auch noch andere Methoden an, die hier nur angedeutet seien: Vergleich mit einem möglichst ähnlichen *Parallel-Test*, sofern einer mit aufgebaut worden ist[8]; *Halbierung* des Tests nach einem ‹zufälligen› oder systematischen Verfahren und Vergleich der beiden Hälften.

Der Zusammenhang zwischen den je beiden Resultatgruppen wird mathematisch in einem Korrelationskoeffizienten ‹r› ausgedrückt. Dieser kann variieren zwischen -1 und $+1$. $r = +1$ bedeutet vollkommene Übereinstimmung: Zum Beispiel die Resultate im ersten Testdurchgang entsprechen absolut denen des zweiten Durchganges. Der Koeffizient beträgt $r = 0$, wenn überhaupt kein Zusammenhang feststellbar ist, und $r = -1$, wenn sich die zweiten Resultate genau umgekehrt wie die ersten verhalten, das heißt, wenn der beste Schüler des ersten Tests der schlechteste im zweiten ist usw. Praktisch ist

[8] Parallel-Tests sind auch aus andern Gründen wünschbar: sie werden gerne eingesetzt zur Bestätigung oder Korrektur des ersten Resultates.

die Übereinstimmung nie absolut ($r = +1$); im allgemeinen wird für gute Tests ein Zuverlässigkeitskoeffizient von $r \geqq +0{,}90$ gefordert.
Der Zeitpunkt der Zuverlässigkeitsprüfung sollte möglichst früh liegen, damit bei schlechter Zuverlässigkeit gleich Verbesserungen angebracht werden können, bzw. in aussichtslosen Fällen nicht noch mehr unnötige Arbeit geleistet wird. Besonders wenn eine Methode gewählt wird, die die Zuverlässigkeitsberechnung für jede Einzelaufgabe zuläßt, ist es vorteilhaft, sie bereits mit der Testvorform – in der Aufgabenanalyse – durchzuführen, damit sich die Auswahl der Aufgaben für die Endform auch auf diese Daten stützen kann.
Der wohl bekannteste Teil der Testaufbau-Arbeit ist die

Eichung.

Sie besteht darin, daß der fertige Test einer großen Anzahl (meist mehrere Tausend) von Personen vorgelegt wird. Die erreichten Testergebnisse je Individuum ergeben zusammen einen Maßstab, mit dem in der spätern Testanwendung die Einzelresultate verglichen werden können. Die Eichmaße werden normalerweise in einer Prozentrangskala oder in einer Standardskala dargestellt. Ein PR (= *Prozentrang*) = 10 bedeutet dabei, daß von der Eichstichprobe 10 Prozent der Individuen ein gleich gutes oder ein schlechteres Resultat erzielten; ein PR = 70 heißt so viel wie, daß nur 30 Prozent der Mitschüler bessere Leistungen vollbringen als der getestete Schüler. Da die meisten Schüler durchschnittlich befähigt sind und die positiven und negativen Abweichungen mit größerem Abstand von der Mitte immer seltener werden (sogenannte Normalverteilung), ist leicht ersichtlich, daß zwischen den Resultaten $PR_1 = 45$ und $PR_2 = 50$ ein kleinerer effektiver Unterschied besteht als zwischen $PR_3 = 5$ und $PR_4 = 10$. Während bei der Interpretation von Prozenträngen immer diese Einschränkungen angebracht werden müssen, messen die sogenannten *Standardnormen* (T-, z-, Z-, C-Skalen usw.) die absoluten Fähigkeitsgrade, das heißt gleichen Skalenabständen entsprechen gleiche Fähigkeitsunterschiede.[9]

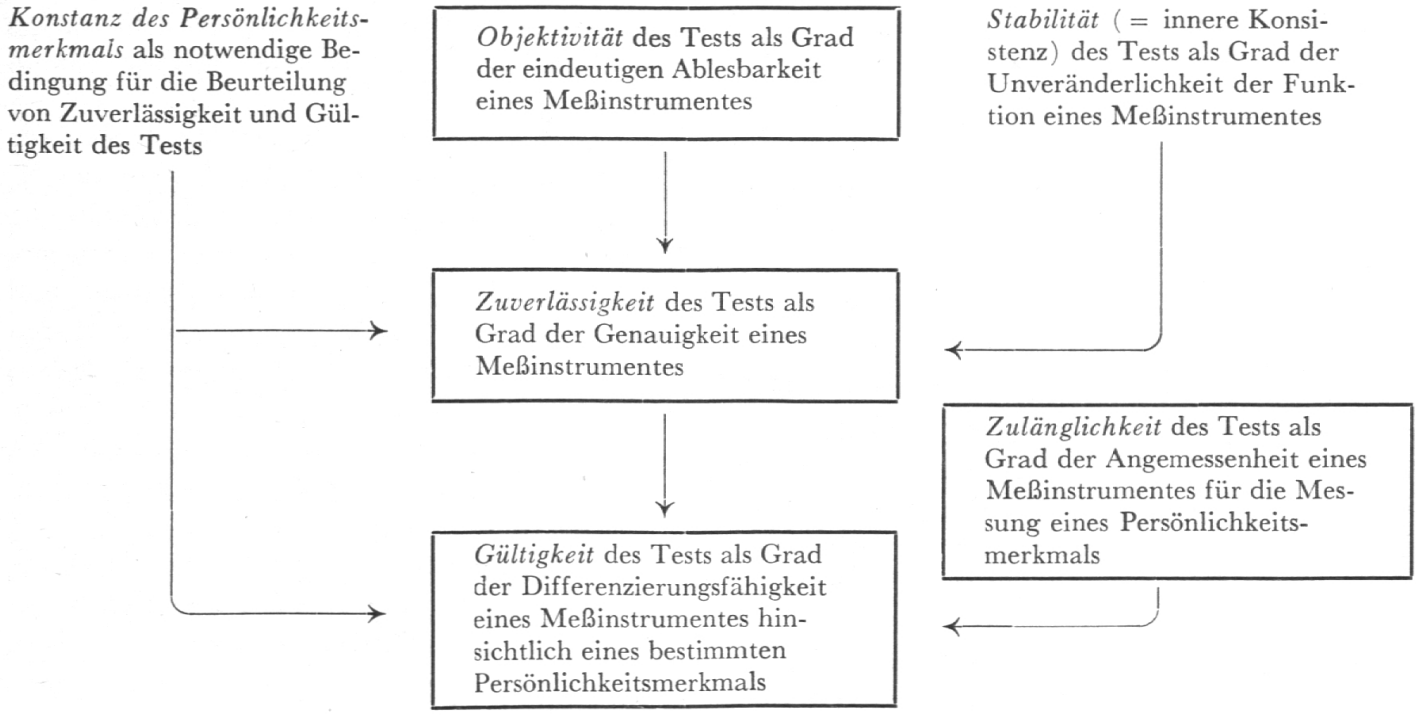
[9] Das komplizierte Skalen-Problem ist damit nur angedeutet. Es ist zu wünschen, daß in den Anleitungen zu Schulleistungstests Interpretationshinweise zur Art der verwendeten Eichskala angebracht werden.

Es ist ganz klar, daß die Eichstichprobe, die gleichsam aus dem Gesamt derjenigen Individuen ausgewählt wird, für die der Test schließlich gedacht ist, für einen verläßlichen Maßstab nicht nur recht groß, sondern auch, wie man sagt, repräsentativ sein muß. *Repräsentanz* bedeutet gleichmäßige Vertretung aller Teile der Gesamtheit. So wären zum Beispiel 2000 Zürcher Sechtskläßler für unser Beispiel eine schlechte Stichprobe, wenn der Test nachher auch für Landkinder und erst noch auf andere Kantone auch anwendbar sein soll.
Schließlich sei noch die wichtigste aller Untersuchungen zum Testaufbau genannt: der Nachweis der

Gültigkeit (englisch validity = Validität).

Während die Zuverlässigkeit eines Tests etwas über das Wie der Messung aussagt, handelt es sich jetzt um das Was. Die Frage nach der Angemessenheit oder *Zulänglichkeit* der Testaufgaben wurde schon am Anfang gestellt und wird nun in präziser Form wiederholt. Wenn immer möglich sind empirische Untersuchungen zu fordern. In unserm Beispiel wären zum Beispiel die spätern Lateinnoten mit den Testresultaten in der sechsten Klasse in Beziehung zu setzen. Es ist meist eine Vielzahl von Kriterien möglich und auch wünschenswert, einerseits zur gegenseitigen Absicherung, andererseits im Hinblick auf eine polyvalente Anwendungsmöglichkeit des Tests. In unserm Fall könnte es zum Beispiel der Testabsicht besser entsprechen, zusätzliche Korrelationen auszurechnen mit einem speziellen Lehrerurteil über die grammatikalischen Fähigkeiten ihrer einzelnen Schüler oder mit einem entsprechenden Test, sofern einer vorliegt. Bedeutungsvoll kann aber schon eine bloße Erfolgskontrolle als Vergleich mit den unmittelbar folgenden Aufnahmeprüfungsresultaten sein.
Die Gültigkeit kann im allgemeinen nur hoch sein, wenn auch die Zuverlässigkeit groß ist, denn wenn die Messung schlecht ist, können Testaufgaben an sich lange *zulänglich*, das heißt dem Ziel angemessen sein, die Gültigkeit leidet mit der Zuverlässigkeit. – Zur bessern Kennzeichnung der wichtigsten Testbegriffe und deren Abhängigkeitsbeziehungen sei hier eine Zusammenstellung nach *Lienert* (1961, S. 17) wiedergegeben:

Fortsetzung Seite 183



So bedeutsam die Gültigkeitskontrolle eines Tests ist, so wichtig ist es auch, zu wissen, daß es recht verschiedene Arten von Gültigkeit gibt, die nicht in jedem Fall gleich angezeigt sind. Die wichtigsten sind:

a) *Die äußere Gültigkeit.* Sie besteht in der Korrelation zwischen den Testresultaten und einem praktischen Kriterium (Bewährung = Gültigkeit auf lange Sicht, zum Beispiel – in unserm Fall – Übereinstimmung mit Maturanoten, Erfolg = gute Korrelation mit einem Kriterium, das relativ rasch nach der Testaufnahme gewonnen wird, zum Beispiel Übereinstimmung mit Aufnahmeprüfungsresultaten). Unter allen Gültigkeitsarten ist diese für den Lehrer die wichtigste.

b) *Die logische Gültigkeit* wird dann ermittelt, wenn es schwer fällt, ein gutes, praktisches Kriterium zu finden. Das trifft vor allem bei sehr spezialisierten Tests zu, bei denen die äußere Gültigkeit eher für eine ganze Batterie zusammen bestimmt werden müßte. Die logische Gültigkeit besteht darin, daß mehrere Sachverständige nach ihrer Ansicht zur Gültigkeit und Angemessenheit des Tests befragt werden. Diese Gültigkeit wird meistens nicht in einer Zahl, sondern in Worten festgehalten.

c) *Die triviale Gültigkeit* liegt dann vor, wenn weder eine Untersuchung noch eine Befragung von Sachverständigen nötig ist, um gewisse Zusammenhänge mit praktischer Sicherheit aussprechen zu dürfen. So kann zum Beispiel die Frage, ob ein Diktat wirklich die Rechtschreibfähigkeit messe, eine triviale sein. Anders verhielte es sich aber, wenn auf Grund eines solchen Rechtschreibtests auf die allgemeine Intelligenz geschlossen werden möchte; diese Fähigkeit des Tests müßte wohl empirisch nachgewiesen werden.

d) *Die innere Gültigkeit* besteht in der Korrelation zu andern Tests, die sich bereits in irgend einer Beziehung als gültig erwiesen haben.

e) *Die faktorielle Gültigkeit* wird durch die moderne mathematische Methode der Faktorenanalyse ermittelt. Sie soll den Test innerhalb eines Faktorenzusammenhangs (von menschlichen Fähigkeiten) festlegen. Im Anschluß an solche sind gezielte Revisionen möglich, die die sogenannte Sättigung mit einem bestimmten Faktor erhöhen und alle andern Faktoren möglichst ausschalten sollen.

Es ist nicht leicht, eine allgemein gültige Unterstgrenze für den Gültigkeitskorrelationskoeffizienten anzugeben. Je nach der Adäquatheit des Kriteriums, dem zeitlichen Abstand und den Erwartungen an einen Test (Beurteilung von individuellen Differenzen oder Gruppendifferenzen) darf der Koeffizient bis auf 0,5, eventuell 0,4 hinunterfallen. Über 0,8 wird er selten steigen.

Teil II und Literaturangaben folgen in der übernächsten Nummer.