ANNELIESE KÜHNECK

Der Test in der Eignungs-Untersuchung

HERAUSGEBER
DEUTSCHE GESELLSCHAFT FÜR PERSONALWESEN
FRANKFURT A·M·

DER TEST

IN DER EIGNUNGSUNTERSUCHUNG

Eine Darstellung der gebräuchlichen Methoden bei der Ausarbeitung

und Bewertung schriftlicher Testverfahren

von

ANNELIESE KÜHNECK

VERLAG KOMMENTATOR G.M.B.H. · FRANKFURT AM MAIN

Druck: Mainzer Verlagsanstalt und Druckerei

V O R W O R T

Die Deutsche Gesellschaft für Personalwesen hat bereits in mehreren Veröffentlichungen ihre Auffassung über die Bedeutung und die Notwendigkeit eines objektiven Prüfungsverfahrens zum Ausdruck gebracht. In der Schrift über „Die Neugestaltung des öffentlichen Dienstes" wurden die Gründe für die Einführung der Methode der sogenannten Eignungsprüfungen in die öffentlichen Verwaltungen dargelegt. Es wurde darauf hingewiesen, daß dieses Verfahren die Fachprüfung nicht ersetzen kann und nicht ersetzen soll, daß diese weiterhin in den meisten Fällen unentbehrlich und eine notwendige Voraussetzung für die Anstellung im öffentlichen Dienst bleibt. Dagegen sollen die Eignungsprüfungen die objektive Auslese gerade dort sichern und erleichtern, wo bisher nur unsichere Prognosen möglich waren. Damit sollten die Eignungsprüfungen zu einem wichtigen, wenn nicht sogar unentbehrlichen Hilfsmittel für die Personalauswahl im öffentlichen Dienst werden, wenn auch die letzte Entscheidung dem Leiter der Behörde vorbehalten bleiben muß.

Die Schrift über „Neuzeitliche Methoden der Personalauslese" befaßt sich mit dem Sinn und Zweck psychologischer Eignungsuntersuchungen und stellt mit allgemeinverständlichen Erläuterungen die einzelnen Verfahren dar, die zur psychologischen Beurteilung des Menschen dienen und zur Anwendung kommen.

Die vorliegende Schrift geht einen Schritt weiter. Sie beabsichtigt, einen Einblick in die Werkstatt für die Schaffung der bei einer objektiven Auslese benötigten Tests zu gewähren. Durch Darstellungen in der Presse und insbesondere auch illustrierten Zeitschriften wird oftmals der Eindruck erweckt, als sei die Abfassung wie die Auswertung dieser Tests eine einfache, wenn nicht sogar unwissenschaftliche Arbeit, und es war nicht selten, daß sich Interessenten an die Deutsche Gesellschaft für Personalwesen wandten, um mit Hilfe von Tests selbst Prüfungen vorzunehmen und Prüfungsergebnisse festzustellen.

Der Leser dieser Schrift wird erkennen, daß es sich bei den Eignungsprüfungen um ein Verfahren nach streng wissenschaftlichen Methoden handelt, das nur in der Hand von Fachleuten wirksam gestaltet werden kann. Nichts wäre für die Einführung dieser Verfahren verhängnisvoller, als wenn diese in der Hand von Laien liegen würden, die glauben, nach der Kenntnis einiger Tests die folgenden selbst ausarbeiten oder gar selbst auswerten zu können.

Es handelt sich bei diesen Prüfungsverfahren noch um eine sehr zarte Pflanze, deren schlechte oder falsche Behandlung dazu führen würde, daß das werdende Vertrauen in die Richtigkeit und Notwendigkeit dieser Verfahren einen Rückschlag erlitte. Dieses Büchlein möge vielmehr dazu beitragen, daß das Vertrauen in diese für viele noch neuen Methoden verstärkt wird. Dann werden Verwaltung und Wirtschaft ihren Vorteil bei der Auswahl des besten Personals haben, und die Bewerber werden erkennen, daß sie auf diese Art den besten Weg zu der Tätigkeit finden, die ihren Anlagen und Fähigkeiten am meisten entspricht und ihnen selbst den Weg für den Beruf ebnet, in dem sie am besten zu eigener Zufriedenheit und für die Allgemeinheit wirken können.

Frankfurt a. Main, im April 1951.

Deutsche Gesellschaft für Personalwesen
Dr. Kurt Oppler

ZUM GELEIT

Endlich setzt sich auch in Deutschland die Überzeugung in weiten Kreisen durch, daß überall dort, wo es sich im Leben um den Menschen handelt — und in allen kulturellen Angelegenheiten steht der Mensch im Mittelpunkt —, die Psychologie als die Wissenschaft von der Eigentümlichkeit und den Spielarten des spezifisch menschlichen, d. h. des körperlich-seelisch-geistigen Geschehens die Grundlage für ein richtiges Handeln bildet, das auf einer richtigen Beurteilung der Tatsachen beruhen muß. Mit dieser Einsicht sind freilich auch die Erwartungen sehr gestiegen, die man an die Psychologie stellt, und sie hat sich ja auch seit fünfzig Jahren in steigendem Maße bemüht, diesen Anforderungen gerecht zu werden.

Dabei ist sie verschiedene Wege gegangen, wie das immer geschieht, wenn ein Erkenntnisgebiet neu erschlossen wird. Ein Teil dieser Wege war unergiebig und führte auch nur zu unsicheren Ergebnissen. Ein Teil hat mehr oder weniger große Fortschritte gebracht. Darunter befindet sich auch der Weg, von dem in dieser Abhandlung die Rede sein soll, der Weg des Testes. Es ist wissenschaftsgeschichtlich interessant zu sehen, wie das Testverfahren, ursprünglich nur als Probe für ganz engbegrenzte Leistungen gedacht, allmählich immer weitere Bereiche des Seelischen und Geistigen in seine Anwendungen einbezieht, indem zugleich die Ergebnisse der allgemeinen psychologischen Erforschung der seelischen Vorgänge verwertet und immer mehr zur Grundlage gemacht werden, und wie dann dieses Testverfahren seinerseits für die allgemeine Psychologie seine Ergebnisse zur Verfügung stellt und so neuen Gesichtspunkten zum Durchbruch verhilft. Die Aufspaltung der Aufmerksamkeit, der Intelligenz, der musikalischen Begabung und mancher anderen psychologischen Vorgänge und Tatbestände in eine Reihe von Faktoren hat das Bild vom Wesen der psychischen Dynamik neu gestaltet und führt uns immer tiefer an den Punkt heran, von dem aus, wie wir hoffen, das menschliche Geschehen in allen seinen Dimensionen durchschaubar wird. Angesichts dieser Entwicklung ist die Haltung derer, die über das Testverfahren geringschätzig hinweggehen wollen, in keiner Weise mehr zu rechtfertigen. Es ist für die Psychologie, und zwar auch für die theoretische Psychologie, als eine der wichtigsten Arbeitsmethoden unentbehrlich geworden.

Die vorliegende Schrift soll dazu dienen, die Grundtatsachen des Testverfahrens einem größeren Kreis von Interessenten bekannt zu machen und ihnen zugleich ein Bild von den Schwierigkeiten zu vermitteln, die beim Aufbau solcher Systeme von Proben zu überwinden sind und ihnen die Sorgfalt und Umsicht zu zeigen, die hier angewandt wird. Dann wird sich ganz von selbst ergeben, daß nur Sachverständige die Tests gebrauchen.

Göttingen, im April 1951. Prof. Dr. J. von Allesch

INHALTSVERZEICHNIS

STATISTISCHER ANHANG

I.

Einleitung

Wahrscheinlich hat noch nie ein wissenschaftliches Prüfverfahren so sehr und so anhaltend das öffentliche Interesse erregt wie der *Test*. Das ist verständlich, wenn man bedenkt, daß sein Gegenstand der Mensch mit seinen Anlagen und Fähigkeiten ist und daß die mit ihm verbundenen Probleme in alle Gebiete des öffentlichen Lebens hineinreichen.

Dem Versuch, mit bestimmten Methoden etwas mehr über einen Menschen zu erfahren, als er sonst nach außenhin zeigt, begegnet man überall. Viele beobachtende und einfallsreiche Menschenkenner (Pädagogen, Politiker, Personalchefs) hatten und haben ihre „Privattests", mit denen sie glauben, ihre Mitmenschen unter die Lupe nehmen und beurteilen zu können. Aber so originell diese „Hausrezepte" oft sind, so sind sie doch kaum dazu angetan, wichtige Fragen der Praxis exakt und zuverlässig zu beantworten.

Darum entstanden im Laufe der Zeit wissenschaftliche Prüfverfahren, die sich überall dort Eingang verschafften, wo man mit der Anleitung und Beurteilung von Menschen zu tun hat. — Arbeitsämter verwenden Tests in der Berufsberatung, Betriebspsychologen bauen Tests in ihre verschiedenen Untersuchungen ein, bei der Auslese von Bewerbern für Schulen, Berufsausbildung oder Neueinstellungen ist der Test ein unentbehrliches Hilfsmittel geworden. Auch in der Erziehungsberatung setzen sich Tests, die gerade auf dem Gebiet der Kinder- und Schulpsychologie sehr exakt ausgearbeitet sind, immer mehr durch.

Neben diesen mit wissenschaftlicher Gründlichkeit unter großem Zeitaufwand ausgearbeiteten Verfahren — eine Vorbereitung dauert im allgemeinen mindestens zwei Jahre — trifft man aber in steigendem Ausmaß auf populär gehaltene „psychologische Tests", deren Entwicklung ihre Hersteller viel weniger Zeit und Mühe gekostet hat, als es dem sich unter Umständen ergebenden materiellen Vorteil entspricht. Es gibt genug Zeitschriften, Illustrierte Zeitungen und Wochenschriften, die zum Käuferfang solche „Pseudo-Tests" abdrucken, aus denen der Leser ohne große Anstrengung die eigene Charakteranalyse, Lebens- und Eheberatung (Prognosen selbstverständlich eingeschlossen!) oder Auskunft über andere wichtige Fragen erhalten soll.

Der unbefangene „psychologische Laie" muß angesichts dieses Durcheinanders verschiedenartiger und verschiedenartiger Tests allmählich ein unbehagliches Gefühl bekommen, besonders wenn er meint, daß von dem zufälligen Ausgang solcher Verfahren unter Umständen die Existenz eines Menschen abhängen könnte. Er vermag ja nicht die „psychologischen Spielereien" von den wissenschaftlich erarbeiteten und praktisch bewährten Tests zu trennen, da ihm nicht bekannt ist, welche Voraussetzungen ein Verfahren erfüllen muß, um wirklich ein *Test* zu sein. Je nach Einsicht und Humor wird er entweder das Kind mit dem Bade ausschütten und die Anwendung von Tests bei entscheidenden Personalfragen grundsätzlich ablehnen, oder er wird sich zunächst über die Möglichkeiten und Grenzen psychologischer Untersuchungsverfahren informieren wollen, ehe er ein endgültiges Urteil fällt.

An solche interessierten und kritischen Nicht-Fachleute richtet sich diese Schrift. Sie will ihnen einen Eindruck vermitteln von der Entstehung und Bewertung wissenschaftlicher Testverfahren, soweit sie im Rahmen von Eignungsuntersuchungen zur Anwendung kommen. Sie beschränkt sich also einerseits auf solche Tests, die die *berufliche Eignung* eines Menschen erfassen, und zum anderen auf die technische Seite dieser Verfahren. *Dementsprechend darf der Leser keine auch nur annähernd erschöpfende Aufzählung oder Beschreibung psychologischer Tests und ihrer gutachtlichen Möglichkeiten erwarten.* Die inhaltliche Seite der Tests muß im Rahmen dieses Themas weitgehend zurücktreten.

Ebensowenig ist es der Zweck dieser Arbeit, ausführlich über den Sinn und die Notwendigkeit von *Eignungsprüfungen* zu sprechen[1]). Anschaulicher als theoretische Erörterungen mögen zwei konkrete Situationen zeigen, vor welche Aufgaben ein Eignungsprüfer gestellt sein kann:

Irgendwo in der Industrie oder in der Verwaltung sollen Nachwuchskräfte eingestellt werden. Ihre Ausbildung bedeutet für den Betrieb oder die Behörde einen erheblichen Aufwand an Zeit und Mitteln. Man ist darum interessiert, möglichst nur solche Bewerber einzustellen, die den beruflichen Anforderungen, welche später an sie gestellt werden, voll gerecht werden können, die also nicht nur Interesse, sondern auch eine ausreichende Begabung für ihre zukünftige Tätigkeit mitbringen.

Die Zahl der Bewerber ist groß, sie unterscheiden sich hinsichtlich Alter und Vorbildung, so daß Schul- und andere Zeugnisse keine sichere Auslesegrundlage bilden können. Alle Bewerber persönlich kennenzulernen und sich dadurch ein Bild von ihren Fähigkeiten und Entwicklungsmöglichkeiten zu machen, ist für den Personalbearbeiter recht schwierig, wenn nicht unmöglich. Auch weiß jeder, der mit der Beurteilung von Menschen zu tun hat, daß der persönliche Eindruck in diesen Dingen sehr unzuverlässig ist und daß man sein Urteil oft auf Grund der später gezeigten Leistungen revidieren muß.

In diesem Fall ist die wissenschaftliche Eignungsuntersuchung ein gutes und auf längere Sicht sehr rentables Hilfsmittel, um eine sachgerechte Auswahl zu treffen. Vor allem wird sie nicht durch die zufällige Zusammensetzung einer Bewerbergruppe beeinflußt, sondern richtet sich in ihrer Bewertung nach feststehenden Maßstäben, die jederzeit nachweisbar und vergleichbar sind.

Ein anderes Beispiel:

Ein junger Mann, der durch ungünstige Zeitumstände einige Jahre der Ausbildung verloren hat, muß sich für einen Beruf entscheiden. Welcher Beruf eignet sich aber am besten für ihn, in welcher Laufbahn hat er seinen Anlagen entsprechend die besten Leistungsmöglichkeiten? — Er hat vielseitige Interessen, war immer ein recht guter Schüler, ohne aber auf einem Gebiet eine besonders deutliche Begabung gezeigt zu haben. Es ist zu überlegen, ob sich das große finanzielle Opfer eines Studiums jetzt noch lohnt oder ob er nicht vielleicht in einem kaufmännischen Beruf mehr leisten würde.
Auch hier könnte eine Eignungsuntersuchung, verbunden mit einer persönlichen Beratung, wahrscheinlich weiterhelfen.

Eignungsprüfungen haben also zweierlei Sinn:

1. *Auslese* bei Neueinstellungen, Beförderungen und Umbesetzungen im Auftrag einer Behörde oder eines Betriebes, und
2. *Beratung* des Einzelnen zu seiner menschlichen oder beruflichen Förderung.

In vielen Fällen wird beides verbunden sein, denn besonders weniger geeignete Bewerber bedürfen ja des Rates und der Hilfe, um ein Arbeitsgebiet zu finden, das ihren Fähigkeiten entspricht.

Es wird Praktiker geben, die einwenden, daß sie auf Grund ihrer Erfahrung und Menschenkenntnis eine Auslese oder Beratung ohne den Aufwand einer Eignungsprüfung ebensogut, vielleicht sogar besser durchführen könnten. Dies entspricht aber keineswegs den Tatsachen, denn es ist in zahlreichen wissenschaftlichen Untersuchungen nachgewiesen worden, daß niemand die Fähigkeit hat, intuitiv oder auch auf Grund eines persönlichen Gesprächs Charaktereigenschaften und Fähigkeiten richtig einzuschätzen[2]). Die Urteile von bewährten und erfahrenen Praktikern über ein und dieselbe Person stimmten, wenn sie ohne objektive Hilfsmittel (Zeugnisse o. ä.) gewonnen waren, selten oder nie überein; oft war es sogar unmöglich zu erkennen, daß es sich dabei um den gleichen Menschen handelte. Der Grund liegt darin, daß jede persönliche Beurteilung notwendig subjektiv gefärbt ist.

Um diesen Mangel bei der Personalauslese und -beratung zu beheben, hat man sich in jahrelanger Arbeit bemüht, Prüfmethoden auszuarbeiten, die eine *objektive* Bewertung der beruflichen Leistungsmöglichkeiten ergeben. Sie haben gegenüber den subjektiven Urteilen den Vorteil einer größeren Genauigkeit, der Nachweisbarkeit und Vergleichbarkeit ihrer Ergebnisse und schließlich einer nicht unerheblichen Zeitersparnis. Viele Fragen, die sonst erst aus langer Zusammenarbeit geklärt werden können, sind durch objektive Prüfmethoden (Tests) unmittelbar zu beantworten. Tests werden daher bei einer modernen Eignungsbegutachtung so weit wie irgend möglich herangezogen. Ihre Ergebnisse bilden das Fundament jeder Beurteilung — wenn sie auch immer durch freiere Verfahren (z. B. Einzelaussprache und Rundgespräche, die *keine Tests* im Sinne dieser Arbeit sind) gestützt und interpretiert werden müssen.

[1]) Hierzu siehe TURK / DORRHOFER: »Neuzeitliche Methoden der Personalauslese.«

[2]) Eine sehr anschauliche Darstellung dieser Probleme findet sich bei WILDE: »Die Frage der Sicherheit in der psychologischen Diagnose I.«

Damit die Tests zu exakten und zuverlässigen Ergebnissen führen — und das ist erforderlich, wenn der Wert einer Eignungsprüfung nicht fragwürdig sein soll —, müssen sie inhaltlich und technisch genau durchgearbeitet sein. Zur Sicherung der technischen Voraussetzungen bedient man sich einer Reihe von Methoden, die auf Ergebnissen der modernen Statistik aufbauen.

Wenn im Verlauf dieser Schrift statistisch-mathematisches Material gebracht wird, so mag jemand, dem es zu fachlich scheint, ruhig darüber hinweglesen. Es ist aber jedem, der sich eingehender mit der Materie beschäftigen will, auch ohne Spezialkenntnisse zugänglich. — Fachleuten sei gesagt, daß diese Arbeit weder ein Lehrbuch noch ein Praktikum, sondern nur ein sehr vereinfachter Überblick sein soll.

II.

Wesen, Arten und historische Entwicklung von Tests

Der von Francis GALTON um 1880 in die Psychologie eingeführte Testbegriff umfaßt jede Prüfung oder Stichprobe, die von bestimmten, genau kontrollierbaren Bedingungen ausgeht und zu einem meßbaren Ergebnis führt. Es hat sich gezeigt, daß man mit richtig angesetzten Stichproben zu Ergebnissen kommt, die nicht nur für die *einzelnen* Verhaltensformen eines Menschen zutreffen, sondern die auch für sein *allgemeines* Verhalten charakteristisch sind.

Je nach der Art und dem Zweck eines Tests stehen die qualitativen *(wesensmäßigen)* oder die quantitativen *(mengenmäßigen)* Ergebnisse im Vordergrund. In jedem Fall müssen jedoch bestimmte Anforderungen erfüllt sein, wenn der Test keine Zufallsergebnisse liefern soll. Die Frage nach diesen Anforderungen oder Wesensmerkmalen ist die Kernfrage bei einem Prüfverfahren überhaupt; erst wenn sie geklärt ist, kann man von einem wirklichen Test sprechen.

A.

Wesensmerkmale

Man muß von einem wissenschaftlichen Testverfahren verlangen, daß es vier wesentliche Voraussetzungen erfüllt: *Gültigkeit, Zuverlässigkeit, Objektivität* und *Normen.*

Über diese Merkmale soll jetzt im einzelnen gesprochen werden.

1. Gültigkeit

Man muß wissen, *was* ein Test erfaßt, denn der beste Test nützt uns nichts, wenn wir nicht sicher sind, ob er wirklich das prüft, was wir damit prüfen wollen. Dieses Wissen kann nicht durch Überlegungen am Schreibtisch gewonnen werden, sondern muß immer durch die praktischen Ergebnisse gesichert sein — wenn auch der theoretische Ansatz bei der Entwicklung

eines Tests nicht unterschätzt werden darf. Aber selbst ein Fachmann kann sich bei Dingen irren, die ganz eindeutig erscheinen mögen. Es genügt darum nicht, auf Grund theoretischer Erwägungen Aufgaben zusammenzustellen, die z. B. die Aufmerksamkeit eines Menschen prüfen sollen, und einen solchen Test als „Aufmerksamkeitstest" zu bezeichnen. Ein anderer Prüfer könnte einen Aufmerksamkeitstest ausarbeiten, dessen Ergebnisse sich von denen des ersten erheblich unterscheiden. Erst der Vergleich mit den im täglichen Leben gezeigten Aufmerksamkeitsäußerungen entscheidet, welcher Test den größeren praktischen Wert hat.

Es ist also notwendig, die Gültigkeit eines Tests zu prüfen, bevor man ihn bei der Begutachtung verwendet. Hierzu müssen die Testergebnisse mit anderen Verhaltensformen oder Leistungen der Prüflinge verglichen werden, die auf die gleiche Funktion zurückzuführen sind. Für einen solchen Vergleich gibt es verschiedene Möglichkeiten, z. B. die Gegenüberstellung von Testresultaten und meßbaren oder schätzbaren fachlichen Leistungen.

Die Aufgabe besteht dabei vor allem darin, einen möglichst zuverlässigen Vergleichsmaßstab zu finden. Hierfür sind objektive Leistungskriterien wie Schulleistungen, Examensergebnisse, Beförderungen, Produktionsmenge, Absatz u. ä. am besten geeignet. Liegen solche nicht vor, so muß man subjektive Beurteilungen durch Lehrer, Meister, Vorgesetzte oder sonstige Fachleute heranziehen. Beispiel:

Wenn eine Gruppe von Verwaltungsschülern einen Denkrechentest bearbeitet hat, so kann man — um die Gültigkeit zu kontrollieren — die Resultate mit den Prüfungsergebnissen der mathematischen Fächer oder mit den Urteilen durch die zuständigen Lehrkräfte vergleichen. Stimmen beide Bewertungen hinreichend überein, so ist das ein Zeichen für die Brauchbarkeit des Tests.

Allerdings sind gerade die subjektiven Beurteilungen selten sehr sicher und genau[1]). Vor allem ist es dabei oft nicht möglich, von den Beurteilern eine differenzierte Einstufung (etwa in 5—10 Klassen) zu verlangen. In diesem Fall kann man u. U. einen anderen Weg einschlagen, nämlich den Vergleich der Testleistungen beruflich tüchtiger und weniger tüchtiger Arbeitskräfte. Beispiel:

Eine Prüfstelle hat mehrere Tests zur Prüfung des Handgeschicks entworfen, das in den verschiedensten industriellen Berufen erforderlich ist. Die Gültigkeit dieser Tests soll untersucht werden, d. h. es soll festgestellt werden, ob und mit welcher Sicherheit sie wirklich das Handgeschick messen. — Zu diesem Zweck setzt sich die Prüfstelle mit einem Industriewerk in Verbindung und läßt von zuständigen Meistern 100 Arbeiter bzw. Arbeiterinnen auswählen, die ausgesprochen geschickt sind, und 100 andere, deren Handgeschick nur mäßig ist. — Diese beiden Gruppen werden mit den neu ausgearbeiteten Verfahren geprüft. Wenn sich in einem Test die Leistungen beider Gruppen nicht oder nur wenig voneinander abheben, so ist das ein Zeichen für die geringe Gültigkeit des betreffenden Verfahrens. Wenn dagegen ein Test eindeutig zwischen den Leistungen beider Gruppen differenziert, ist er zur Prüfung des Handgeschicks geeignet.

In ähnlicher Weise läßt sich die Gültigkeit eines Tests kontrollieren, der bestimmte *berufstypische* Fähigkeiten erfaßt, man vergleicht die Leistungen zweier deutlich verschiedener Berufsgruppen. Beispiel:

Ein Test zur Prüfung der mechanischen Sorgfalt wird einer größeren Gruppe bewährter Schreibkräfte (Sekretärinnen und Stenotypistinnen) gegeben. Zum Vergleich zieht man eine Gruppe von Angehörigen solcher Berufe heran, bei denen mechanische Sorgfalt keine wesentliche Rolle spielt. — Auch hier müssen sich die Testleistungen beider Gruppen deutlich unterscheiden, wenn der Test für die Prüfung dieser speziellen Fähigkeit gültig sein soll. Ein derartiger Vergleich ergab bei einem Arbeitstest, dem später zu beschreibenden Durchstreich-Test, eine deutliche Überlegenheit der Schreibkräfte gegenüber anderen Gruppen des Verwaltungsdienstes.

Eine ergiebigere, wenn auch sehr langwierige Methode der Gültigkeitskontrolle ist der Vergleich von Testergebnissen mit bestimmten fachlichen Leistungen über eine längere Zeitspanne hin. Beispiel:

[1]) Vgl. MEILI: »Psychologische Diagnostik« S. 11 ff.

Die Testprüfung wird am Beginn der Berufsausbildung durchgeführt, zu einem Zeitpunkt also, an dem noch nichts über die Begabung oder die fachlichen Leistungen der Prüflinge bekannt ist. Mindestens ein Jahr nach Abschluß der Ausbildung, wenn bereits Aussagen über die fachliche Tüchtigkeit gemacht werden können, werden die Leistungen dieser Prüflinge von zuständigen Beurteilern eingeschätzt. (Die Testergebnisse dürfen diesen Beurteilern nicht bekannt sein.) Wenn nötig, kann man die Beurteilung in bestimmten Zeitabständen wiederholen lassen.

Auf diese Weise hat man die Möglichkeit, die berufliche Entwicklung zu verfolgen und die Testergebnisse damit zu vergleichen, d. h. man kann feststellen, ob ein Test eine Begabung prüft, die für die betreffende fachliche Leistung erforderlich ist und die vielleicht erst im Verlauf einer längeren Ausbildung oder Berufsausübung hervortritt. Wenn ein Test schon *vor* Beginn der Ausbildung zu ähnlichen Ergebnissen führt wie die späteren fachlichen Beurteilungen, so ist er zur Vorhersage der fachlichen Leistungsfähigkeit geeignet und kann bei entsprechenden Eignungsprüfungen angewandt werden.

Die hier angeführten Beispiele zeigen nur einige der möglichen Wege, die Gültigkeit eines Tests zu untersuchen. Ihnen gemeinsam ist die Notwendigkeit, einen geeigneten und möglichst zuverlässigen *Vergleichsmaßstab* für die Kontrolle der Testergebnisse zu finden. Da diese Maßstäbe nicht immer zuverlässig sind, müssen in der Regel mehrere Wege eingeschlagen werden, bevor der Grad der Gültigkeit eines Verfahrens bestimmt werden kann. Über die technischen Einzelheiten der entsprechenden Methoden wird im Anhang zu dieser Arbeit zu sprechen sein (Korrelationsrechnung).

Die Probleme sind schon an dieser Stelle so eingehend erörtert worden, um dem Leser deutlich zu machen, daß die zentrale Frage der Gültigkeit eines Tests ohne das Verständnis und die Mitarbeit zuständiger Fachleute kaum geklärt werden kann.

2. Zuverlässigkeit

Man muß nicht nur wissen, *was* ein Test mißt, sondern auch *wie genau* er mißt. Denn die Ergebnisse haben nur wenig Wert, wenn sie von äußeren Fehlern oder von Konstruktionsmängeln des Tests sehr stark beeinflußt werden. Ein solcher Test wäre etwa zu vergleichen mit einem Zollstock, der sich durch Witterungseinflüsse so stark dehnt oder zusammenzieht, daß er bei wiederholten Messungen des gleichen Gegenstandes deutlich verschiedene Resultate zeigt. Ebenso wie jedes physikalische Meßinstrument muß auch ein Test bei wiederholten Prüfungen zu gleichen Ergebnissen führen. Von geringen zufälligen Schwankungen kann dabei, dem Gegenstand psychischer Messungen entsprechend, abgesehen werden.

Die beste Methode, um die Zuverlässigkeit eines Tests zu kontrollieren, ist die *Testwiederholung*. Sie darf nicht in einem zu großen Zeitabstand erfolgen, da sich sonst die im Prüfling liegenden Voraussetzungen verändert haben könnten (besonders bei solchen Verfahren, die sich auf das Wissen eines Menschen beziehen). — Es ist allerdings oft nicht möglich, einen Test mehrere Male mit den gleichen Prüflingen durchzuführen, denn bei vielen Verfahren können sich die Prüflinge noch ihrer früheren Lösungen erinnern und die Aufgaben dann viel sicherer und schneller erledigen als beim ersten Male. In diesen Fällen müssen andere Methoden zur Kontrolle der Zuverlässigkeit angewandt werden.

Wenn ein *Paralleltest* besteht, d. i. ein Verfahren, welches in Inhalt, Schwierigkeit und Form dem ersten völlig entspricht, ohne genau die gleichen Einzelaufgaben zu enthalten, kann man die Ergebnisse beider Verfahren miteinander vergleichen und dadurch den Grad der Zuverlässigkeit feststellen.

Eine dritte Möglichkeit der Zuverlässigkeitskontrolle ist die Methode der *Testhalbierung*, bei der zwei sich entsprechende Teile des gleichen Tests verglichen werden. Wenn beide Teile einen Menschen in derselben Weise charakterisieren, so ist das ein Zeichen dafür, daß die Teilergebnisse nicht dem Zufall unterliegen.

Auf die technische Seite dieser Methoden wird noch einzugehen sein. Sie führen zu einem zahlenmäßigen Ergebnis, einem Korrelationskoeffizienten, der ausdrückt, in welchem Grade die Resultate beider Messungen übereinstimmen bzw. wie groß die Fehler sind, die bei der Entstehung dieser Resultate mitgewirkt haben.

Diese Fehler können in der Konstruktion des Tests selbst liegen und müssen dann auf dem Wege einer eingehenden Analyse der Testergebnisse behoben werden. (Wenn das nicht möglich ist, kann man einen Test nur mit großem Vorbehalt anwenden.) Sie können auch auf Faktoren außerhalb des Tests beruhen, die immer in verschiedenem Maße mitwirken und sich nicht völlig beseitigen lassen (sog. zufällige Fehler). Hierher gehören Aufmerksamkeitsschwankungen, Stimmungs- und Gefühlsmomente, Übungsfähigkeit und andere beim Prüfling liegende Ursachen, außerdem Störungen bei der Durchführung des Tests, Auswertungsfehler usw. Die letzte Gruppe ist unvermeidlich, wenn bei der Durchführung und Auswertung eines Tests subjektive Momente mitspielen.

3. Objektivität

Es ist notwendig, daß ein Test die *objektive Bewertung* der einzelnen Leistungen ermöglicht. Die Beurteilung darf also nicht durch subjektive, im Prüfer oder Auswerter liegende Fehler oder durch unzuverlässige Bewertungsmethoden beeinflußt werden. Dafür müssen folgende Voraussetzungen erfüllt sein:

a) Ein Test muß — unabhängig vom Zeitpunkt und Ort der Prüfung — immer in genau der gleichen Weise *durchgeführt* werden, so daß für alle Prüflinge die gleichen Bedingungen bestehen. Dazu ist es vor allem erforderlich, daß jedesmal wörtlich dieselbe Instruktion (möglichst gedruckt) gegeben und die Arbeitszeit genau eingehalten wird.

b) Die *Auswertung* der Testunterlagen muß frei von willkürlichen Entscheidungen (z. B. Sympathien und Antipathien) des Auswerters sein. Sie erfolgt deshalb in der Regel anonym, d. h. die nur mit einer Prüfnummer versehenen Testunterlagen werden von Auswertern bearbeitet, denen die Prüflinge unbekannt sind oder die zumindest nicht wissen, wie sich die Nummern verteilen. Außerdem kontrollieren sich jeweils zwei oder mehr Auswerter gegenseitig.

c) Die *Bewertung* der Testleistungen muß nach objektiven Richtlinien erfolgen, die dem Aufbau und der Eigenart des Tests angepaßt sind.

4. Normen

Ein Test muß genau feststehende Bewertungsmaßstäbe oder *Normen* haben, da man sonst nichts über den Wert der festgestellten Leistungen aussagen kann. Zunächst ein praktisches Beispiel:

Wenn verschiedene Meister gebeten werden, Aussagen über das Arbeitstempo bestimmter Arbeiter zu machen, die sie eine Zeitlang bei ihrer Tätigkeit beobachten konnten, so werden ihre Urteile recht verschieden ausfallen, denn jeder Beurteiler hat — je nach den Anforderungen, die er an sich selbst und an andere stellt und je nachdem, in welchem Arbeitskreis er lebt — andere Maßstäbe. Es ist also durchaus möglich, daß der gleiche Mann von einem Meister als ein sehr schneller, von einem anderen als ein nur durchschnittlicher Arbeiter bezeichnet wird. Außerdem kann das Urteil durch die Zusammensetzung der beobachteten Gruppe beeinflußt werden. Besteht sie vorwiegend aus guten Kräften, wird der Maßstab unwillkürlich nach oben verschoben, so daß eine durchschnittlich schnelle Arbeitskraft hier wahrscheinlich negativ beurteilt würde.

In einer ähnlichen Situation ist man bei der Durchführung eines Tests, der keine Normen hat. Man kann einem einzelnen Ergebnis nicht ansehen, ob es gut oder schlecht ist, solange man es nur mit den Resultaten einer kleinen Prüfgruppe vergleichen kann. Denn es ist für die

Beurteilung einer Testleistung erforderlich, die Leistungen einer großen Zahl von Angehörigen derjenigen Berufs- und Altersklassen zu kennen, bei denen der Test angewandt wird. Erst wenn man weiß, wie sich ihre Ergebnisse verteilen, kann eine Testleistung objektiv bewertet werden.

Es gibt verschiedene Möglichkeiten, die Höhe eines Testresultates zu charakterisieren. Häufig werden — wie es auch in den Schulen üblich ist — die Leistungen mit Prädikaten bzw. Ziffern bezeichnet. Z. B. sehr gut (1) — gut (2) — befriedigend (3) — ausreichend (4) — ungenügend (5).

Bei allen derartigen Einteilungen muß man jedoch wissen, wie sich die Leistungen sämtlicher, der Eichung zugrunde liegender Prüflinge auf die einzelnen Wertgruppen verteilen, ob etwa jedes Prädikat gleich viele Leistungen umfaßt oder ob die mittleren Gruppen dichter belegt sind als die extremen. — Es ist z. B. ein Unterschied, ob die „sehr guten" Leistungen 5, 10 oder 20 % aller Resultate ausmachen. Bei der Einteilung in wenige Wertgruppen können darüber hinaus die Leistungen, die auf der Grenze zwischen zwei Gruppen liegen, u. U. ungerecht beurteilt werden.

Um solche Unklarheiten zu vermeiden und außerdem eine genauere Bewertung der Leistungen zu ermöglichen, ist es üblich, die einzelnen Testresultate als *Prozentränge* auszudrücken.

Ein Prozentrang gibt an, wieviel Prozent der Gruppe, an der der Test geeicht worden ist, in ihren Leistungen unter dem betreffenden Wert liegen. — Der Prozentrang 100 (P_{100}) würde also bedeuten, daß der Prüfling der beste der gesamten Gruppe ist; der Prozentrang 0 (P_0) würde besagen, daß der Prüfling der schlechteste der ganzen Gruppe ist. Bei P_{50} läge die Hälfte aller Leistungen über, die andere unter dem betreffenden Wert. — Prozentränge zeigen also die *Stellung der Prüflinge* innerhalb ihrer Berufsgruppe und nicht das Quantum der Leistung, etwa die Zahl der richtig gelösten Aufgaben. Die Prozentränge eines Prüflings in verschiedenen Tests können unmittelbar verglichen und — zur besseren Übersicht für den Gutachter — in einem Leistungsprofil graphisch dargestellt werden. Das Prinzip der Prozentränge, über deren Berechnungsmethode noch zu sprechen sein wird, hat sich in der Praxis weitgehend durchgesetzt.

Eine noch exaktere Art von Normwerten, die neben den Prozenträngen immer mehr gebräuchlich werden, sind die *„Standardwerte"*. Sie sind vor allem für differenzierte statistische Berechnungen geeignet. Auch die Standardskala wird aus der Verteilung sämtlicher Testergebnisse gewonnen. Ihre Stufen entsprechen aber nicht — wie die der Prozentskala — gleich großen Gruppen von Prüflingen, sondern gleich großen *Leistungsunterschieden*. Dabei wird jedes Testresultat seinem relativen Abstand vom Mittelwert (Durchschnitt) entsprechend ausgedrückt.

Auch die Standardwerte gestatten den Vergleich der verschiedenen Testleistungen eines Prüflings bzw. der Leistungen mehrerer Prüflinge. Sie ermöglichen darüber hinaus eine differenzierte Kombination von Testwerten, so daß man mit ihrer Hilfe die Ergebnisse eines Prüflings in verschiedenen Tests mathematisch zusammenfassen kann. Wenn in einem Eignungsbefund — das ist das schriftliche Endurteil einer Eignungsprüfung — die einzelnen Leistungsgebiete auch immer getrennt dargestellt werden müssen, um ein Bild von der individuellen Leistungspersönlichkeit zu geben, so ist es *daneben* manchmal zweckmäßig, für jeden Prüfling einen solchen Gesamtwert anzugeben.

Die technischen Einzelheiten der hier nur erst angedeuteten Fragen sollen — da sie die Kenntnis bestimmter statistischer Grundbegriffe voraussetzen — im Anhang besprochen werden.

Eine unerläßliche Voraussetzung für die Festsetzung von Normen ist die richtige Zusammensetzung der Prüfgruppe, an der diese Normen gewonnen werden. Man muß bei der Auswahl der Prüflinge ebenso vorgehen wie in allen Fällen, in denen aus der Untersuchung oder Befragung eines Bevölkerungsausschnittes Schlüsse auf die Gesamtbevölkerung oder auf eine

bestimmte Bevölkerungsschicht gezogen werden. Dazu ist es notwendig, daß man eine möglichst große und charakteristische Gruppe von Menschen aus denjenigen Schichten erfaßt, für die der betreffende Test gültig sein soll.

Dementsprechend gelten Normen meist auch nur für bestimmte Gruppen von Menschen und können nicht beliebig auf andere Gruppen übertragen werden. — Wenn z. B. ein Test für Verwaltungsbeamte der mittleren Laufbahn geeicht worden ist, kann man die Normen nicht auf Fabrikarbeiter, Studenten oder Schulkinder übertragen.

Ein guter Test hat in der Regel *verschiedene* Normen für die einzelnen Alters-, Bildungs- oder Berufsgruppen. Dadurch kann jeder Prüfling mit dem ihm entsprechenden Maßstab gemessen werden.

Die hier beschriebenen Wesensmerkmale, Gültigkeit, Zuverlässigkeit, Objektivität und Normen bilden die Voraussetzung für den Wert eines Tests. Man muß vor der Anwendung jedes Verfahrens — sei es nun selbst konstruiert oder übernommen — zuerst in einer Probeprüfung untersuchen, ob diese Merkmale zutreffen.

Mit ihnen verbunden sind bestimmte praktische Forderungen, die man in folgende zwei Punkte zusammenfassen kann:

1. Ein Test soll ein Minimum an Zeit, Kosten und Mühe in der Durchführung, Auswertung und Deutung erfordern.

Man muß einen Test also so konstruieren, daß er mit möglichst geringer zeitlicher und kräftemäßiger Belastung für Prüfling und Prüfer und mit möglichst wenig Materialaufwand zu guten und zuverlässigen Ergebnissen führt.

2. Ein Test soll dem geistigen Niveau der Prüflinge in Stoff und Schwierigkeit angepaßt sein.

Er soll also — soweit das ohne Beeinträchtigung seines Wertes möglich ist — solche Aufgaben oder Probleme enthalten, die für den Prüfling interessant sind, und die ihm nicht so fern liegen, daß er mit Unlust oder Nachlässigkeit an die Lösung herangeht. Das ist selbstverständlich nur in gewissen Grenzen möglich. Denn man will auch das Verhalten gegenüber eintönigen und wenig anziehenden Arbeiten erfassen; außerdem würde die Einschränkung auf ein bestimmtes Interessengebiet einen Teil der Prüflinge dem anderen gegenüber bevorzugen (besonders bei einem Test, der nicht nur für eine engbegrenzte Laufbahn angewandt werden soll). — Darüber hinaus besteht bei zu großer Berufsnähe eines Verfahrens immer die Gefahr, daß Übung und Erfahrung die Resultate verschieden stark beeinflussen.

Die Anpassung der Schwierigkeit eines Tests an das Niveau der Prüflinge ist deshalb zu wünschen, weil die Nachlässigkeit bei zu leichten Aufgaben oder die Entmutigung bei zu schweren dadurch vermieden werden.

Bevor in den folgenden Abschnitten (III und IV) auf die technischen Methoden der Konstruktion und Eichung von Tests eingegangen wird, sei zunächst noch eine Einteilung der gebräuchlichen Testverfahren und ein kurzer Überblick über ihre historische Entwicklung gegeben.

B.

Einteilung der Tests

Man kann Tests nach verschiedenen Gesichtspunkten einteilen: nach dem *Gebiet*, das sie erfassen, nach dem *Material*, aus dem sie aufgebaut sind, nach der *Form* ihres Aufbaues, nach der Art ihrer *Durchführung*, der Methode ihrer *Bewertung* usw. — Es wäre zu verwirrend,

hier alle diese Einteilungsmöglichkeiten zu besprechen; im Verlauf dieser Arbeit wird an den entsprechenden Stellen — soweit es zum Verständnis erforderlich ist — darauf eingegangen werden.

Hier seien zwei dieser Einteilungsmöglichkeiten herausgegriffen, da man an ihnen deutlich machen kann, mit welcher Art von Tests man es in Eignungsprüfungen vor allem zu tun hat.

1. Je nachdem, ob ein Testverfahren den Persönlichkeits-, Begabungs- oder Leistungsbereich erfassen will, kann man drei Gruppen von Tests unterscheiden:

a) Persönlichkeitstests (charakterologische Tests),

b) Begabungs- oder Fähigkeitstests,

c) Leistungstests zur Prüfung fachlichen Wissens und Könnens.

Bei den *Persönlichkeitstests* liegt der Akzent auf dem Erfassen der Struktur der Gesamtpersönlichkeit und ihrer Gestaltungsmöglichkeiten. Bei einem Teil dieser Verfahren überträgt der Prüfling seine charakteristischen Verhaltensweisen und Einstellungen auf die im Test gegebenen Bedingungen. Diese Tests werden darum auch „projektive" *(übertragende)* Verfahren genannt. — Persönlichkeitstests im engeren Sinne erfassen die Züge, die im Verhalten der Umwelt gegenüber (z. B. Gefühlsstabilität, mitmenschliches Empfinden, Willensstetigkeit) wichtig sind.

Die *Begabungs- oder Fähigkeitstests* heben bestimmte mehr oder weniger komplexe Bereiche der Begabung heraus (Bereiche der Intelligenz, der Aufmerksamkeit, des Gedächtnisses, des Arbeitsverhaltens usw.). — Hierher gehören Tests zur Prüfung der allgemeinen geistigen Begabung und solche zur Erfassung spezieller Fähigkeiten (technisches Verständnis, Handgeschick, Reaktionsgeschwindigkeit usw.). — Sie zeigen, welche Anlagen zum Ausüben oder Erlernen bestimmter Tätigkeiten vorhanden sind; von den Testleistungen kann nicht nur auf die augenblickliche Leistungsfähigkeit eines Prüflings, sondern auch auf seine Entwicklungsmöglichkeiten in den entsprechenden Bereichen geschlossen werden.

Die *fachlichen Leistungstests* prüfen die *Kenntnisse* und Fertigkeiten, über die ein Mensch zum Zeitpunkt der Eignungsuntersuchung im Hinblick auf bestimmte berufliche Anforderungen verfügt. Ihr Stoff muß in Zusammenarbeit mit Fachleuten, die für die einzelnen Leistungsgebiete kompetent sind, ausgewählt werden. Sie gehen insofern über die üblichen Examina oder Fachprüfungen hinaus, als für alle Prüflinge die gleichen exakt ausgearbeiteten Bedingungen und Bewertungsmethoden geschaffen sind, die Beurteilung also objektiviert ist.

Jede Eignungsprüfung setzt sich aus Persönlichkeits- und Fähigkeitstests zusammen, während fachliche Leistungstests nur in bestimmten Fällen herangezogen werden (z. B. zur Fachprüfung von Schreibkräften).

Bei Eignungsprüfungen, die zum Zwecke einer individuellen *Beratung* durchgeführt werden, können die Persönlichkeitstests — jedoch nur auf Wunsch des zu Beratenden — erweitert werden. Sie berühren dann persönliche Bezirke, die für eine gewöhnliche Eignungsprüfung nicht von Belang sind.

Bei *Auslese*prüfungen dagegen, in denen die Begabung für bestimmte Berufe oder Leistungsgebiete festgestellt werden soll, liegt der Schwerpunkt auf den Fähigkeitstests. Jedoch sind auch hier zur Interpretation der Ergebnisse Persönlichkeitstests unentbehrlich, denn die festgestellten Begabungen müssen immer von der Struktur der Gesamtpersönlichkeit her gesehen werden.

Wenn sich die vorliegende Arbeit auf Fähigkeitstests beschränkt, so geschieht das, weil man an ihnen am besten die technischen Methoden der Entwicklung und quantitativen Bewertung von Testverfahren demonstrieren kann. Es soll damit keinesfalls der Eindruck erweckt werden, daß den Fähigkeitstests im Rahmen einer Eignungsprüfung eine größere Bedeutung zukommt als den anderen hierbei üblichen Methoden.

2. Nach einem anderen mehr äußerlichen Einteilungsgesichtspunkt kann man Einzeltests (individuelle Tests) und Gruppentests unterscheiden.

Der *Einzeltest*, bei dem der Begutachter nur einem oder einigen wenigen Prüflingen gegenübersteht, hat den Vorteil einer unmittelbaren und freieren Gestaltung der Lösungen, erlaubt die Klärung von Mißverständnissen und das persönliche Eingehen auf den einzelnen Prüfling. Er ermöglicht dem Begutachter die Beobachtung des Verhaltens und der Arbeitsweise und erleichtert ihm dadurch die Deutung der Ergebnisse. Außerdem kann, wenn es erforderlich ist, kompliziertes Material (Apparate u. dgl.) gebraucht werden.

Der *Gruppentest*, mit dem eine größere Anzahl Personen gleichzeitig geprüft werden kann, hat den Vorteil einer großen Zeit- und Kostenersparnis in Durchführung und Auswertung. Er sichert im allgemeinen eine objektivere Bewertung und damit eine genauere Beurteilung der einzelnen Testleistungen, die an den allgemeinen Normen bestimmter Berufs-, Alters- oder Bildungsgruppen gemessen werden.

Einzeltests, die in der Regel mündlich durchgeführt werden, können bei größeren Auslese-prüfungen nur in Ausnahmefällen herangezogen werden, während die schriftlich durchgeführten Gruppentests einen wesentlichen Bestandteil dieser Eignungsprüfungen ausmachen.

C.

Historische Entwicklung von Gruppentests

Die theoretische Psychologie hat seit dem Ende des 19. Jahrhunderts — angefangen mit den Arbeiten von WUNDT (Leipzig) und seinem Schüler CATTELL (USA.) — eine große Zahl von Forschungsexperimenten ausgearbeitet und durchgeführt, um allgemeinpsychologische Probleme sowie speziellere Fragestellungen zu untersuchen und damit wissenschaftliche Hypothesen zu unterbauen.

Einige dieser Forschungsexperimente konnten von der angewandten Psychologie, die sich mit der Lösung praktisch-psychologischer Probleme beschäftigt, übernommen und zu Testverfahren umgearbeitet werden. Daneben wurden neue, den Anforderungen der Praxis angepaßte Methoden entwickelt, die — zunächst als Einzeltests — bei Berufsberatungen, Lehrlingsprüfungen, Begabtenauslesen usw. durchgeführt wurden. Hier ist vor allem die Testskala von BINET und SIMON (1905) zu nennen, deren Aufgabe es war, in den Schulen solche Kinder herauszufinden, die infolge mangelnder geistiger Begabung den Anforderungen einer Volksschule nicht gewachsen waren. Diese Testskala ist — sehr erweitert und mehrfach überarbeitet — auch heute noch eine der führenden Methoden in der Kinderpsychologie.

Da Einzeltests nicht für alle praktischen Aufgaben geeignet waren, wurden allmählich auch Gruppentests entwickelt, mit denen man eine große Anzahl von Menschen gleichzeitig prüfen konnte. Zu ihrer Durcharbeitung und zur Bewertung ihrer Resultate bediente man sich in steigendem Ausmaß statistischer Methoden, die bereits Francis GALTON am Ende des vorigen Jahrhunderts zur Interpretation von Testergebnissen herangezogen hatte.

Die ersten Gruppenuntersuchungen großen Stils wurden 1917/18 in der amerikanischen Armee durchgeführt, die alle neueingestellten Soldaten (etwa 1,75 Millionen) auf allgemeine geistige Fähigkeiten hin prüfte. Seitdem wurden in den Vereinigten Staaten im Zusammenhang mit der allgemeinen Forderung nach Wirtschaftlichkeit und Leistungssteigerung Gruppentests bei Neueinstellungen in Verwaltung und Wirtschaft weitgehend angewandt. Sie bestanden zunächst vor allem aus Intelligenztests, zu denen allmählich auch Tests für spezielle Fähigkeiten und Leistungen hinzukamen.

Die Testmethode hatte dort vor allem bei Schul- und Personalfragen bald einen sehr großen Erfolg, weil sie sich auf allen Gebieten des täglichen Lebens als außerordentlich praktisch erwies. Außer den Army-Tests, mit denen die Eignung für bestimmte militärische Laufbahnen festgestellt wird, wurden dort Serien von Tests für die Anstellung im Verwaltungsdienst, in der Industrie, für den Eintritt in höhere Schulen, Universitäten usw. eingeführt.

Neben der Entwicklung neuer Verfahren kam es zu einer ständigen Verbesserung der statistischen Methoden und der rein technischen Seite der Tests, wie es auf dem europäischen Kontinent wegen der Andersartigkeit der Aufgaben- und Problemstellung sowie aus materiellen Gründen in dem Maße nicht möglich war.

Auch in England sah man schon früh die Notwendigkeit psychologischer Prüfverfahren, besonders bei der Auslese von Bewerbern des Civil Service (auch der höheren und höchsten Laufbahnen). Hierbei wurde allerdings nicht so sehr die Feststellung der momentanen Leistungsfähigkeit für bestimmte Stellungen angestrebt (wie in USA.), sondern vielmehr die Auslese allgemein begabter und entwicklungsfähiger Nachwuchskräfte. Inzwischen hat auch das amerikanische System diesen zweiten Gesichtspunkt der Eignungsbegutachtung stärker berücksichtigt.

Im kontinentalen Europa führten einige Stellen — z. B. die Armeen mehrerer Staaten — zwar ebenfalls Gruppenprüfungen durch, doch spielte die exakte zahlenmäßige Bewertung der Testleistungen dabei keine so wesentliche Rolle wie in den amerikanischen und englischen Tests[1]). In gewissem Umfang wurden Gruppenprüfungen vor allem in Schulen und größeren Betrieben durchgeführt, deren Methoden und Ergebnisse in der Regel aber auf den Arbeitskreis des begutachtenden Institutes beschränkt blieben (z. B. hatten nach dem ersten Weltkrieg in Deutschland Firmen wie AEG, Borsig, Krupp, MAN u. a. ihre eigenen Prüfstellen).

In Deutschland war durch die Jahre der Isolierung der Kontakt mit den wissenschaftlichen Arbeiten des Auslandes verlorengegangen, so daß der Austausch von Ergebnissen und Erfahrungen auf dem Gebiet der angewandten Psychologie nicht möglich war. Und selbst wenn der Kontakt jetzt allmählich wiederhergestellt wird, so kann man doch nicht die Methoden und Ergebnisse anderer Länder einfach übernehmen und auf unsere Verhältnisse übertragen. Neben den Schwierigkeiten der genauen Übersetzung (jedes veränderte Wort schafft unter Umständen andere Bedingungen) haben auch die ausländischen Normen für unsere Bevölkerung in der Regel keine Gültigkeit.

Es ist also notwendig, für neue praktische Aufgaben auch neue Verfahren zu entwickeln bzw. alte umzuarbeiten und sie durch zahlreiche Probeprüfungen in eine Form zu bringen, die den Anforderungen eines wissenschaftlichen Tests (Wesensmerkmale) gerecht wird.

Im allgemeinen macht man sich keine Vorstellung davon, was alles mit der Entwicklung eines Testverfahrens verbunden ist. Um einen Überblick zu geben, sollen hier folgende Fragen erörtert werden:

1. Wie entstehen Gruppentests?
2. Wie werden die einzelnen Testleistungen bewertet?
3. Mit welchen Methoden prüft man die Zuverlässigkeit und Gültigkeit eines Tests?

[1]) Über die Bedeutung der deutschen Wehrmachtspsychologie siehe ANSPACHER: »Bleibendes und Vergängliches aus der deutschen Wehrmachtspsychologie.«

III.

Konstruktion von Gruppentests

Nehmen wir an, eine Prüfstelle habe die Aufgabe, die Eignungsprüfung einer großen Zahl von Bewerbern für den mittleren Verwaltungsdienst vorzubereiten und eine Reihe geeigneter Tests dafür zusammenzustellen. Dabei sind folgende Punkte zu bedenken:

1. Die *Berufsbilder* der entsprechenden Laufbahnen müssen bekannt sein. Ein Berufsbild ist die „eingehende Schilderung einer bestimmten Berufstätigkeit, die das Wesen, die Entwicklung, die Arbeitsaufgabe und -beschreibung, die seelischen und körperlichen Anforderungen ... umfaßt“[1]). Vor allem muß man also die Anforderungen kennen, die an Begabungen und Fähigkeiten der Prüflinge zu stellen sind. Denn der Eignungsgutachter muß Aussagen machen können, ob die Anlagen mit den Anforderungen des Berufes in Einklang stehen und für welche Laufbahnen sich die einzelnen Bewerber eignen.

2. Die Bewerber haben, da sie noch *vor* der Ausbildung stehen, weder fachliches Wissen noch praktische Erfahrung im Verwaltungsdienst. Die Prüfung kann also nicht in der Feststellung der augenblicklichen fachlichen Tüchtigkeit der Bewerber bestehen, sondern muß sich auf die Untersuchung ihrer Begabungen und Entwicklungsmöglichkeiten richten. Es werden darum vor allem *Fähigkeitstests* zur Anwendung kommen.

3. Da es sich um eine größere Zahl von Bewerbern handelt, die in möglichst kurzer Zeit begutachtet werden sollen, wird man im wesentlichen *Gruppentests* heranziehen müssen.

Unter diesen Voraussetzungen müssen die verschiedensten Verfahren zusammengestellt werden. Ihre Zahl muß zunächst möglichst groß sein, damit man im Verlauf von Vorprüfungen diejenigen Tests herausfinden kann, welche sich für die vorliegende Aufgabe am besten eignen. Darunter müssen sich einige befinden, die sich auf die Prüfung der allgemeinen intellektuellen Begabung beziehen und einige, die sich auf andere Leistungsbereiche z. B. auf das Arbeitsverhalten eines Prüflings richten.

Es ist eine allgemein bekannte Tatsache, die schon in der Schule zu beobachten ist, daß überdurchschnittliche Intelligenz allein nicht ausreicht, um gute Leistungen zu erzielen, und daß ein fleißiger, ausdauernder und sorgfältiger Arbeiter mit geringerer Intelligenz unter Umständen mehr leistet als ein intelligenterer, dem es an Arbeitsbereitschaft und Willenssteuerung fehlt. Jedoch muß in fast allen Berufen ein bestimmtes *Mindestmaß* an Auffassungs- und Denkvermögen vorhanden sein, besonders wenn es sich — wie in unserem Fall — um die Prüfung von neu anzulernenden Kräften handelt. Dementsprechend können, wie wissenschaftliche Untersuchungen gezeigt haben, auf Grund einer reinen *Intelligenzprüfung* mit größerer Sicherheit die für eine bestimmte Aufgabe *ungeeigneten* Bewerber festgestellt werden, während für eine zuverlässige Beurteilung der positiven Leistungsfähigkeit immer auch einige andere Tests (z. B. Arbeitstests) herangezogen werden müssen.

Im folgenden soll je ein Test zur Prüfung der Intelligenz und des Arbeitsverhaltens herausgegriffen werden, um an diesen Beispielen die technischen Methoden der Entwicklung von Fähigkeitstests darzustellen.

[1]) SCHARMANN - DÜRRHÜFER: »Berufsbilder aus Verwaltung und Wirtschaft.«

A.

Konstruktion eines Tests zur Erfassung intellektueller Fähigkeiten

Zur Prüfung der intellektuellen Fähigkeiten muß man eine Reihe von Tests durchführen, unter denen der *Analogietest* häufig zur Anwendung kommt. Darin sind einzelne Aufgaben gegeben, bei denen der Prüfling die innere Beziehung zwischen zwei (oder mehr) Begriffen erkennen und auf Begriffe anderen Inhalts übertragen soll. (Die untenstehenden Beispiele werden die Aufgabe verdeutlichen.) Der Prozeß kann sich sowohl an sprachlichem wie an sprachfreiem Material vollziehen. So gibt es Wortanalogien, Figurenanalogien, Bildanalogien. Der einfacheren Darstellbarkeit halber sind für unser Beispiel Wortanalogien gewählt.

1. Zusammenstellung von Einzelaufgaben und Instruktion

Bei der Zusammensetzung eines solchen Tests geht man von einer großen Anzahl Einzelaufgaben aus, die nach der Art der in ihnen enthaltenen Beziehungen (kausal, temporal, modal usw.) und nach ihrem sachlichen Inhalt weitgehend verschieden sind. Außerdem ist darauf zu achten, daß die Lösungen der Einzelaufgaben von speziellen Kenntnissen oder vom Bildungsniveau weitgehend unabhängig sind, da man eine Fähigkeit und kein Wissen prüfen will. Man wählt also Aufgaben wie:

gut	:	schlecht	=	groß	:	*(klein)*
Himmel	:	blau	=	Gras	:	*(grün)*
Fisch	:	schwimmen	=	Hund	:	*(laufen)*
Krankenhaus	:	Patient	=	Gefängnis	:	*(Häftling)*
				(usw. im ganzen 40—50 Aufgaben)		

Sie werden unter Fortlassen des letzten Begriffes ihrer wahrscheinlichen Schwierigkeit nach angeordnet.

Sehr wichtig ist es, daß die *Instruktion* (Arbeitsanweisung) für einen Test so einfach und klar wie möglich formuliert und immer in der gleichen Weise gegeben wird (Objektivität der Durchführung!). Diese Forderung wird bei schriftlichen Anweisungen, die dem eigentlichen Test mit einigen Übungsbeispielen auf einem besonderen Blatt vorangestellt werden, am besten erfüllt. Es hat sich gezeigt, daß bei mündlicher Instruktion die Resultate und die durchschnittliche Arbeitszeit der einzelnen Prüfungsgruppen oft recht unterschiedlich waren.

Die Instruktion für den Analogietest müßte etwa lauten:

„In der jetzt folgenden Aufgabe werden Sie in jeder Zeile drei Wörter sehen. Für das vierte Wort ist ein freier Platz gelassen. Dieses vierte Wort sollen Sie selbst einsetzen. Es muß zum dritten Wort die gleiche innere Beziehung haben wie das zweite Wort zum ersten.

Z. B.: hell : dunkel = groß : *(klein)*

Da „dunkel“ der Gegensatz von „hell“ ist, muß das gesuchte Wort der Gegensatz zu „groß“ sein. — Es müßte also das Wort „klein“ eingesetzt werden.

Die innere Beziehung zwischen den beiden Worten ist nicht immer ein Gegensatz.

Weizen : Brot = Trauben : *(Wein)*

Hier ist „Wein“ die Lösung, denn aus Weizen kann Brot gemacht werden, aus Trauben Wein. Setzen Sie die folgenden Lösungen selbst ein:

Gehorsam	:	Lob	=	Ungehorsam	:	.....
Hahn	:	krähen	=	Hund	:	.....
Forscher	:	Wissenschaft	=	Maler	:	.....

Bitte wenden Sie das Blatt erst, wenn der Prüfleiter das Zeichen dazu gegeben hat."

Sollte sich in Vorversuchen zeigen, daß die Instruktion in dieser Form nicht verstanden wird, so muß sie entsprechend abgeändert werden.

2. *Auswahl der Prüflinge („Probanden")*

Es wurde bereits erwähnt, daß es für die Konstruktion und Eichung eines Tests sehr wichtig ist, eine große und nach bestimmten Gesichtspunkten ausgewählte Gruppe von Menschen zur Verfügung zu haben. Diese Menschen sollen ja diejenige Bevölkerungsschicht vertreten *(repräsentieren)*, für die der Test später angewandt werden soll. Man spricht darum von einer „repräsentativen Stichprobe" aus der betreffenden Bevölkerungsschicht.

Bei der Entwicklung von Kindertests, die für bestimmte Altersstufen geeicht werden, ist es nicht schwierig, solche Stichproben zu gewinnen, da man durch die Schulen an *alle* Kinder zwischen 6 und 15 Jahren herankommen kann. Durch die Erfassung gleichaltriger Klassen aus verschiedenen Schulen (Volks-, Mittel-, höhere Schule) aus verschiedenen Orten (Stadt, Land) und verschiedenen Gegenden ist die Möglichkeit gegeben, einen guten Ausschnitt aus jeder Altersstufe zu erhalten.

Die Gewinnung der Stichprobe ist schwieriger, wenn es sich um die Prüfung von Jugendlichen oder Erwachsenen handelt. Aber da man bei diesen Altersstufen in der Regel einen Test nur für bestimmte Bevölkerungsschichten, etwa für bestimmte Berufsgruppen anwendet, ist es meist nicht nötig, eine Stichprobe aus der Gesamtbevölkerung zu gewinnen. Wenn — wie in unserem Beispiel — Bewerber für den mittleren Verwaltungsdienst geprüft werden sollen, könnte man etwa sämtliche Beamten der mittleren Laufbahn aus verschiedenen Behörden zu einer Probeprüfung heranziehen. Das ist in der Regel ohne Störung des Dienstbetriebes nicht möglich. Darum ist es einfacher, alle Teilnehmer von Verwaltungslehrgängen (für Sekretäre und Inspektoren) zu erfassen, die zu einem bestimmten Zeitpunkt an verschiedenen Verwaltungsschulen ausgebildet werden. In einem Schulbetrieb ist es leichter, gelegentlich einige Stunden für solche Probeprüfungen frei zu machen. Das geht selbstverständlich nicht ohne das Interesse und die Hilfe des Lehrkörpers und der Lehrgangsteilnehmer selbst.

Die Zahl der Probanden ist für die ersten Vorversuche weniger wichtig als eine gute Auswahl. 50 bis 100 Personen genügen in der Regel, um einen ersten Eindruck von der Brauchbarkeit und den Verbesserungsmöglichkeiten eines Tests zu geben. Die *Eichung* eines in mehreren Probeprüfungen überarbeiteten Verfahrens erfordert dagegen mindestens 200 Angehörige der betreffenden Berufsgruppe.

Es ist dabei nicht zulässig, von sogenannten „ausgelesenen Gruppen" auszugehen, etwa nur Menschen, die sich freiwillig zur Verfügung stellen, Angestellte *einer* Behörde, Teilnehmer eines Lehrgangs heranzuziehen. Jeder Praktiker weiß, daß solche Gruppen charakteristische Besonderheiten zeigen können, die dann den Aufbau und die Normen des neuen Tests beeinflussen würden. Ganz falsch wäre es, die Probeprüfungen an einer völlig anderen Bevölkerungsschicht durchzuführen — den für Verwaltungsangestellte gedachten Test etwa an Schülern, Studenten, Sozial- und Industriearbeitern zu entwickeln —, da sich dadurch die Normen erheblich verschieben könnten.

Die hier für die Auswahl der Probanden angedeuteten Regeln gelten nicht nur für die Konstruktion und Eichung eines Tests, sondern ebenso für alle Fälle, in denen man Aussagen

über bestimmte große Gruppen (verschiedene Berufe, Altersstufen, männliche und weibliche Bevölkerung) machen will. Immer müssen die zugrunde liegenden Stichproben genügend groß und repräsentativ sein.

Die *Güte* einer Stichprobe kann man durch den Vergleich zweier kleinerer Stichproben oder durch die Gegenüberstellung der Ergebnisse mit entsprechenden Resultaten anderer Prüfstellen kontrollieren. Wenn etwa eine Prüfstelle mit bestimmten Testmethoden erhebliche Unterschiede zwischen den Ergebnissen männlicher und weiblicher Probanden festgestellt hat, die bei einer anderen Prüfstelle viel weniger deutlich oder gar nicht vorhanden waren, so muß der Grund für die mangelnde Übereinstimmung zunächst in der Zusammensetzung der erfaßten Gruppen gesucht werden. Er kann auch in der Untersuchungsmethode selbst liegen, die sich jedoch — wie wir noch sehen werden — mit sehr differenzierten technischen Mitteln überprüfen läßt.

3. *Durchführung der Probeprüfung*

Um die praktische Brauchbarkeit eines auf Grund theoretischer Erwägungen konstruierten Tests (oder einer Testreihe) zu untersuchen und ihn zu einem zuverlässigen Meßinstrument zu entwickeln, bedarf es wiederholter Probeprüfungen. Die einzelnen Prüfgruppen dürfen dabei zunächst nicht zu groß sein, da sonst eine exakte Durchführung des Tests nicht möglich ist (höchstens 20 Personen).

Es ist wichtig, daß die Probanden über den Sinn solcher Verfahren allgemein und über den Zweck einer Probeprüfung im besonderen unterrichtet werden, damit sie nicht das Gefühl haben, auf eine indirekte Weise im Auftrag der Schule oder Behörde „überprüft" zu werden, andererseits aber diese Vorversuche nicht zu leicht nehmen und sich dementsprechend zu wenig einsetzen. Nun vermag es ein erfahrener und geschickter Prüfer durchaus, die richtige Arbeitseinstellung bei der Gruppe hervorzurufen — auch bei anfänglichen Schwierigkeiten konnten ein steigendes Interesse und zunehmender Eifer während solcher Probeprüfungen beobachtet werden. Man muß allerdings mit den technischen Methoden der Ausarbeitung von Tests vertraut sein, um die statistische Bedeutung eventueller Besonderheiten einer Prüfgruppe richtig beurteilen zu können.

Für jeden Test wird zunächst die gedruckte Instruktion (s. S. 21) gegeben. Der Prüfer überzeugt sich, ob sie von allen Probanden verstanden worden ist, um eventuelle Unklarheiten zu beseitigen und sie in einer späteren Fassung ausschalten zu können. Bei mündlichen Instruktionen, die sich bei manchen Tests zunächst nicht umgehen lassen, oder bei zusätzlichen Erläuterungen muß er genau wissen, was er dabei sagen darf, und was er *nicht* sagen darf. Denn jedes einzelne Wort kann unter Umständen die Prüfbedingungen verändern. Es erübrigt sich zu betonen, daß und warum eine Unterhaltung oder eine Zusammenarbeit während der Lösung des Tests untersagt werden muß.

Wenn alle Probanden wissen, welche Aufgabe sie bei dem vorliegenden Test zu erfüllen haben, wird gleichzeitig mit der Bearbeitung begonnen. Nach der Bearbeitung aller Aufgaben läßt jeder Proband seine Arbeitszeit notieren. Dieses Festhalten der einzelnen Zeiten ist bei großen Gruppenprüfungen wegen der damit verbundenen Unruhe kaum durchzuführen, also müssen die Tests so durchgearbeitet werden, daß sie später mit begrenzten Arbeitszeiten gegeben werden können. Bei der Probeprüfung jedoch ist es notwendig, daß jeder Proband sämtliche Einzelaufgaben bearbeitet, und daß der Prüfer darüber hinaus feststellen kann, wieviel Zeit im Durchschnitt für die Lösung des gesamten Tests benötigt wird.

Die in der Probeprüfung gewonnenen Resultate werden dann einer eingehenden Analyse unterzogen. Sie dient dazu, die Schwierigkeit und Eindeutigkeit der einzelnen Aufgaben festzustellen und die für den Test weniger geeigneten Aufgaben auszuschalten.

4. Analyse und Auswahl der Einzelaufgaben

Wenn im weiteren Verlauf dieser Arbeit von „Aufgaben" gesprochen wird, so sind damit die Einzelaufgaben eines Tests gemeint. Jeder gute Test besteht aus einer größeren Anzahl solcher Einzelaufgaben.

Zunächst muß für jede Aufgabe festgestellt werden, wieviel Prozent der Probanden sie richtig, falsch oder gar nicht gelöst haben. Dabei zeigen sich in der Regel erhebliche Unterschiede, deren Ursachen zu untersuchen sind.

Man wird vielleicht Aufgaben finden, die von mehr als 10 Prozent der Probanden gar nicht gelöst wurden. Bei ihnen sind entweder die Ausgangsbegriffe zu wenig bekannt, um eine innere Beziehung abheben zu können (Bildungsmoment), oder die Relation als solche ist logisch nicht eindeutig, so daß nur unexakte Lösungen möglich sind. In diese Gruppe gehören Analogien wie:

Ursache	:	Wirkung	=	Grund	:	*Folge*
Jähzorn	:	Haß	=	Augenblick	:	*Dauer* / *Zeit* / *Weile*

Man kann oft nicht voraussehen, wie die wirklichen Lösungen ausfallen werden oder welche Schwierigkeiten sie enthalten.

Weiter gibt es etwa Aufgaben, bei denen verschiedene mehr oder weniger richtige Lösungen vorgekommen sind. Der Auswerter greift in diesem Falle manchmal zu dem Ausweg, die nicht ganz exakten Lösungen als „halbrichtig" zu bewerten. Zeigt eine Aufgabe viele solcher halbrichtigen Lösungen, so ist sie unsachgemäß formuliert. Z. B.:

Mutter	:	Kind	=	Baum	:	*Frucht* / *Sprößling* / *Setzling* / *Ableger*
Glauben	:	Erkennen	=	Kirche	:	*Wissenschaft* / *Universität* / *Forschungsinstitut* / *usw.*

Man sieht schon in der Anlage dieser Aufgaben, daß sich die Beziehung der beiden ersten Wörter nicht eindeutig übertragen läßt, streng genommen also keine Analogie vorliegt.

Eine dritte, ebenfalls nicht geeignete Art von Aufgaben sind die, welche von mehr als 90 Prozent aller Probanden richtig gelöst worden sind, die also zu leicht sind und damit keinen diagnostischen Wert haben. Z. B.:

Speicher	:	Weizen	=	Bücherschrank	:	.....
gut	:	schlecht	=	Recht	:	.....

Es ist zweckmäßig, ein bis zwei dieser leichten Aufgaben an den Anfang des Tests zu stellen oder sie — mit einigen schwereren — als Übungsbeispiele der Instruktion anzufügen. Die übrigen zu leichten Aufgaben müssen — ebenso wie die vorher angeführten Beispiele (bildungsgebundene, mehrdeutige und unexakte Aufgaben) — ausgeschaltet werden.

Alle Aufgaben, die bei der Analyse ihrer Lösungen keine groben Mängel gezeigt haben, können in die erste Fassung des Tetsts übernommen werden. Die bei ihnen aufgetretenen Fehllösungen lassen Fehler oder Ungenauigkeiten im Denken erkennen. Z. B.:

						falsch	*richtig*
Kreis	:	Kugel	=	Quadrat	:	*„Rechteck"*	*Würfel*
Roman	:	Leser	=	Film	:	*„Schauspieler"*	*Kinobesucher*
Haus	:	Hütte	=	Fluß	:	*„Strom"*	*Bach*

(Im letzten Beispiel ist die Beziehung umgekehrt worden; ein Fehler, der nicht selten vorkommt.)

Wenn nach der ersten Analyse nicht mehr genug Aufgaben zur Verfügung stehen, muß die Probeprüfung mit einer erweiterten Fassung des Tests an einer anderen Gruppe von Probanden wiederholt werden.

Auf die Möglichkeit einer eingehenderen statistischen Analyse der Einzelaufgaben soll hier nur hingewiesen werden. Sie wird durchgeführt, um die „innere Stabilität" eines Tests zu prüfen, d. h. um festzustellen, ob jede einzelne Aufgabe diejenige Funktion erfaßt, die man mit dem gesamten Test messen will, oder ob etwa bestimmte Aufgaben aus dem Rahmen der übrigen herausfallen. Hierfür werden sämtliche Probanden, an denen der Test erprobt worden ist, nach einem Gültigkeitsmerkmal z. B. nach der Höhe ihrer Leistungen in der Verwaltungsschule (s. S. 12) in drei oder vier Gruppen eingeteilt. Für jede einzelne Aufgabe wird festgestellt, von wieviel Prozent jeder dieser Gruppen sie richtig gelöst worden ist. Diejenigen Aufgaben zeigen die größte Gültigkeit, die einen deutlichen Leistungsabfall von der besten zur schlechtesten Gruppe erkennen lassen. — Im Interesse der Verkürzung und Präzisierung eines Tests sollte man für die endgültige Fassung möglichst nur solche Aufgaben zusammenstellen, die einen hohen Gültigkeitsgrad haben.

5. Aufbau des Tests

Hat man festgestellt, welche Aufgaben für den Test geeignet sind, so kann mit dem Aufbau begonnen werden. Hierfür gibt es drei Möglichkeiten:

1. Man stellt Aufgaben von annähernd gleicher Schwierigkeit zusammen, und zwar möglichst nur solche, die von etwa 50 Prozent sämtlicher Probanden richtig gelöst worden sind. Dabei ist es wichtig, daß die Anzahl der Aufgaben nicht zu gering ist, damit der Test — wenn er mit begrenzter Arbeitszeit gegeben wird — genügend verschiedene Wertmöglichkeiten zuläßt (30 bis 40 Aufgaben). — Diese Form eignet sich vor allem für größere Gruppenprüfungen bei Probanden von einigermaßen gleichartigem geistigem Niveau, also besonders für Ausleseprüfungen.

2. Man ordnet die Aufgaben nach gleichmäßig ansteigender Schwierigkeit, also etwa: Aufgabe 1 mit 95 Prozent, Aufgabe 2 mit 92 Prozent, Aufgabe 3 mit 89 Prozent richtigen Lösungen usw. bis etwa 10 Prozent. Dabei soll die durchschnittliche Schwierigkeit aller Aufgaben etwa bei 50 Prozent liegen. — Diese Form ist für die Prüfung einer sehr wenig einheitlichen Gruppe geeignet oder für solche Fälle, in denen die absolute Begabungshöhe eines Prüflings beurteilt werden soll, also vor allem für Beratungen.

3. Eine dritte Art des Testaufbaues ist die Aufstellung von „Spiraltests", in denen Einzelaufgaben verschiedener gut konstruierter Tests gemischt sind. — Diese Form hat den Vorteil, daß die Zeitansagen und Pausen zwischen den einzelnen Tests vermieden werden und daß damit die Prüfung ruhiger und schneller ablaufen kann. (Ein Spiraltest benötigt etwa 20—30 Minuten Arbeitszeit.) Die von manchen Prüfstellen angewandten kombinierten Intelligenztests sind in dieser Form aufgebaut.

Falls sehr viele brauchbare Einzelaufgaben zur Verfügung stehen, kann ein „Paralleltest" aufgestellt werden, also ein zweiter Test, der mit dem ersten in Anzahl und Schwierigkeit der Aufgaben übereinstimmt und der ihm auch inhaltlich entspricht. Solche Paralleltests sind in bestimmten Fällen der Praxis sehr wertvoll, z. B. bei Wiederholungsprüfungen, Bekanntwerden einer Fassung usw.

6. Wahllösungen

Eine sehr gute und immer mehr verwendete Konstruktionsmethode ist der Aufbau eines Tests nach dem Prinzip der Wahllösungen. Zwei Beispiele aus dem Analogietest sollen diese Form veranschaulichen:

Krankheit	:	Gesundheit	=	Armut	:	?	
a) Elend		b) Reichtum		c) Arbeit		d) Niedrigkeit	e) Zustand

Vogel	:	Flügel	=	Fisch	:	?	
a) Gräten		b) Wasser		c) Schuppen		d) Flossen	e) Kiemen

Für jede Aufgabe sind also mehrere Lösungen zur Auswahl gestellt, unter denen der Prüfling die richtige zu kennzeichnen hat (wie es in den Beispielen geschehen ist). Die Schwierigkeit der selbständigen sprachlichen Formulierung, die bei den früher angeführten Beispielen zu dem geforderten Denkprozeß hinzukommt, und die Abhängigkeit dieser Formulierungen von den momentanen Einfällen bzw. Hemmungen fallen dabei fort. Außerdem werden Unterschiede in der Schreibgeschwindigkeit ausgeschaltet. — Wahllösungen sind vor allem dann unentbehrlich, wenn man bei den Prüflingen sprachliche Gewandtheit oder Schreibgeschick nicht voraussetzen kann.

Ein weiterer sehr wesentlicher Vorteil dieser Methode ist die größere Objektivität der Auswertung, die besonders bei sprachlichen Tests stark ins Gewicht fällt, da es sonst oft nicht zu vermeiden ist, daß mehrere richtige oder fast richtige Lösungen möglich sind.

Bei dieser Form kann man einen Test dadurch beliebig erschweren, daß man die nicht exakten Lösungen inhaltlich oder formal der richtigen Lösung annähert. Beispiel:

genau	:	ungefähr	=	Sorgfalt	:	?	
a) Schnelligkeit		b) fehlerhaft		c) Bemühung		d) Pedanterie	e) Flüchtigkeit

Fläche	:	Körper	=	Malerei	:	?	
a) Farbe		b) Modell		c) Plastik		d) Bild	e) Tiefendimension

Bei dieser Art der Aufgabenstellung ist die Möglichkeit des unkritischen Ratens nie ganz auszuschalten. Die Chance, daß ein Proband dabei richtig „tippt", ist jedoch um so geringer, je größer die Zahl der Wahllösungen ist. Während man bei zwei bis drei Lösungen die Zufallstreffer durch bestimmte Berechnungen korrigieren muß, bewirkt eine Korrektur bei vier und mehr Lösungen nur sehr geringe Änderungen der Ergebnisse, so daß sie im allgemeinen nicht notwendig ist.

Im Interesse einer objektiven Durchführung und Auswertung sollte man bei größeren Gruppenuntersuchungen, für die nur wenig Zeit zur Verfügung steht, möglichst alle, besonders die nicht ganz eindeutig lösbaren Tests, auf die zuletzt beschriebene Form (vier bis fünf Wahllösungen) umstellen. — Dem Einwand, daß diese Form die Selbständigkeit und Produktivität des Denkens beeinträchtige, kann dadurch begegnet werden, daß neben diesen objektiven Verfahren immer auch einige andere durchgeführt werden, die speziell auf eine produktive Form des Denkens gerichtet sind, und die diesen Faktor dann auch deutlicher in Erscheinung treten lassen.

Ein Analogietest, der bei der Auslese von Bewerbern für den mittleren Verwaltungsdienst angewandt werden soll, muß sich also aus mindestens 30 Aufgaben von annähernd gleicher Schwierigkeit zusammensetzen. Die zuverlässigste Form erhält dieser Test, wenn man ihn nach dem Prinzip der Wahllösungen aufbaut und wenn er nur wirklich gültige Einzelaufgaben enthält, wenn also jede Aufgabe die Prüflinge in der gleichen Weise aufteilt wie der gesamte Test. Diese feinere Durcharbeitung eines Verfahrens kann gewöhnlich erst im Verlauf einer längeren Anwendungszeit erfolgen, nachdem eine sehr große Zahl von Probanden mit dem Test geprüft worden ist. Eindeutigkeit, abgestimmte Schwierigkeit und genügende Anzahl der Einzelaufgaben dagegen sind Forderungen, die schon vor der Eichung eines Tests erfüllt sein müssen, da sonst die objektive Bewertung der Leistungen in Frage gestellt ist.

Ähnlich wie in dem hier dargestellten Beispiel der Analogietests verläuft die Konstruktion aller anderen Verfahren, mit denen die verschiedenen Bereiche der Intelligenz erfaßt werden

sollen. Der Schwerpunkt liegt dabei jeweils auf der Auswahl und Zusammenstellung geeigneter Einzelaufgaben, deren Schwierigkeit so abgewogen sein muß, daß nur ein Teil der Prüflinge sie richtig lösen kann. Denn nur so ist es möglich, Leistungsunterschiede festzustellen und zahlenmäßig auszudrücken.

B.

Konstruktion eines Tests zur Erfassung des Arbeitsverhaltens

Ein Arbeitstest soll Auskunft über den Arbeitsverlauf eines Menschen geben, er soll zeigen, ob ein Prüfling schnell oder langsam, sorgfältig oder flüchtig, gleichmäßig oder sprunghaft, konzentriert oder unaufmerksam, ausdauernd oder unstet arbeitet. Die darin gestellte Aufgabe muß so einfach sein, daß alle Prüflinge sie lösen können. Der Test darf sich nicht aus verschiedenen, getrennt zu lösenden Problemen (Einzelaufgaben) zusammensetzen wie ein Intelligenztest, sondern muß durchgehend die gleiche Tätigkeit erfordern. Denn nur so kann man Unterschiede im Arbeitsverlauf feststellen.

Ein Verfahren, mit dem einige der obengenannten Faktoren erfaßt werden können, ist der altbewährte Durchstreichtest von BOURDON (1895), der als Beispiel angeführt werden soll. Im Durchstreichtest muß der Prüfling Reihen von Buchstaben durchsehen und dabei bestimmte Buchstaben streichen. Diese gleichförmige und uninteressante Arbeit verlangt eine erhebliche Anspannung der Aufmerksamkeit (Konzentration) über einen längeren Zeitraum hin. Die in vielen Berufen — nicht zuletzt bei Bürotätigkeit — vorausgesetzte Konzentrationsfähigkeit kann sehr stark von Willensfaktoren beeinflußt werden.

Der Durchstreichtest ist in verschiedenen Fasssungen gebräuchlich. Einige, wie der Bourdon-Test selbst, verwenden sinnvolle Texte, in denen dann bestimmte Buchstaben gestrichen werden müssen; andere enthalten Gruppen von Buchstaben, die ohne Sinnzusammenhang nebeneinander gestellt sind (nach WHIPPLE); wieder andere bestehen aus Figuren oder Zahlen, um die Lesefähigkeit auszuschalten; eine weitere Art schließlich erfordert das Erfassen und Durchstreichen bestimmter Buchstabenzusammenstellungen (nach STERZINGER).

Für die Prüfung berufstätiger Erwachsener, bei denen man die Vertrautheit im Umgang mit Schriftzeichen voraussetzen kann, haben sich sinnfreie Buchstabengruppen als recht geeignet erwiesen.

Die Konstruktion und Eichung eines solchen Tests erfordert weit mehr psychologische Durcharbeitung als die eines Verfahrens, das auf einen engeren Funktionsbereich gerichtet ist. Es würde über den Rahmen dieser Schrift hinausgehen, die mit dem Durchstreichtest verbundenen p s y c h o l o g i s c h e n Probleme zu erörtern. Darum seien nur sehr vereinfacht einige Faktoren herausgegriffen, die man etwa mit diesem Test erfassen kann:

1. Die Konzentrationsfähigkeit (Intensität der Aufmerksamkeit), die sich in der Menge der in einer bestimmten Zeit richtig gestrichenen Buchstaben zeigt.
2. Die Gleichmäßigkeit bzw. Ungleichmäßigkeit der Aufmerksamkeit, die aus der Verteilung der Fehler zu erkennen ist.
3. Das Arbeitstempo bei vorwiegend mechanischer Tätigkeit, das sich in der Zahl der bearbeiteten Reihen ausdrückt.
4. Die Sorgfalt, erkennbar an dem Verhältnis von richtig gestrichenen zu übersehenen Buchstaben.
5. Das Verhalten neuen Anforderungen gegenüber (Umstellbarkeit), das man durch einen Wechsel der Instruktionen erfassen kann.

6. Die Störbarkeit bei mechanischer Tätigkeit, die sich durch das Einsetzen von Störungen während der Arbeit feststellen läßt.

Es gibt noch andere Ansatzmöglichkeiten für die Anwendung dieses Tests. Sie sollen hier nicht einzeln besprochen werden. Wesentlich ist es, den Test so zu konstruieren. daß man in den Ergebnissen die verschiedenen Seiten der Testleistung klar trennen kann, damit einer willkürlichen Deutung möglichst enge Grenzen gesetzt sind.

1. Zusammenstellung der Buchstabenreihen und Instruktion

Wenn sich auch ein Arbeitstest nach den erfaßten Funktionsbereichen in Inhalt, Aufbau und Bewertungsmöglichkeiten weitgehend von einem Intelligenztest unterscheidet, so gelten doch für seine Konstruktion die gleichen Regeln:

Die Instruktion muß einfach und allgemein verständlich sein; es darf nicht mehrere Lösungsmöglichkeiten geben (Eindeutigkeit). Voraussetzungen an Bildung und Erfahrung müssen auf ein Minimum beschränkt bleiben; der Schwierigkeitsgrad muß gleich oder gleichmäßig ansteigend und dem Niveau der Prüflinge angepaßt sein. Außerdem muß der Test einen gewissen Umfang haben, damit jeder Proband Gelegenheit hat, seine Fähigkeiten zu zeigen.

Diese technischen Regeln sind hier leichter einzuhalten als bei einem Intelligenztest, da durchgehend die gleiche Arbeit (das Durchstreichen bestimmter Buchstaben) geleistet werden soll. Man muß lediglich bei der Zusammenstellung der Buchstabengruppen darauf achten, daß die Zahl der zu streichenden Buchstaben innerhalb der einzelnen Reihen einigermaßen konstant ist und daß ihre Stellung in sinnvoller Weise wechselt. — Wenn man diese Bedingungen darüber hinaus für alle im Test enthaltenen Buchstaben etwa gleichmäßig einhält, kann man später die Instruktion ändern (d. h. andere Buchstaben streichen lassen), ohne ein anderes Testformular zu benötigen. Damit erspart man sich die Konstruktion von Paralleltests, die bei den Intelligenztests so schwierig und zeitraubend ist.

Ein nach diesen Gesichtspunkten zusammengestellter Test sieht etwa folgendermaßen aus (hier sind nur die ersten 8 Reihen wiedergegeben):

byj pkh ypd qtf yfg kpq pft gyq pgb tdy kht fbq hqf jqb
thd gqf qhj bgk bkd jht dpk hjt fkj hgq byg pdj tgd ykp

jhy pdb kbp bgf fyq dft ptg dqj fyp jqh kty gqf kfp tkb
tqp kjt pfy bkd tyk jbq dqt pfh fyq kdb gdb jfq pjt hpq

hyf tpj hyk dqf dfy gph kqp fyp jpq tfh gjp ydh kyd bhp
gpq kdb pbf gtj jhg bkt yjg dkb fkg ydb tqf gkb qtj tfq

dyb ghf hjt qyg bdp ghf bhj kgy dpj gtk pyt hkp bfd ykg
tqp kjt pfy bkd tyk jbq dqt pfh fyq kdb gdb tfq pft hpg

Die Instruktion für den Durchstreichtest würde z. B. lauten:

„In der folgenden Aufgabe sind Ihnen Reihen von Buchstaben gegeben, die ohne sinnvollen Zusammenhang nebeneinander stehen. Sie sollen diese Reihen einzeln — wie beim Lesen eines Buches — durchgehen und dabei bestimmte Buchstaben streichen.

Üben Sie diese Aufgabe zunächst an den folgenden Reihen. Streichen Sie dabei deutlich alle **p** und **t**, wie es in der ersten Reihe bereits geschehen ist!"

dhy ~~pfg~~ d~~p~~h jg~~p~~ gyd ~~pf~~j fy~~p~~ ~~t~~dy f~~p~~j khg yb~~p~~ g~~t~~l fq~~p~~ d~~t~~y
jfb ykq qyb ftk pht kbq ktb gjk qdt ybq hjq kfj htb gjk

hgd pkf fkd hbq gjd tyk ykg bqd phq fpb kfy dhf dtk jkb
pjb ypj yjp gty pbf jqh tjd hbf iyk dgt pqj gbt fqg byh

„Verfahren Sie — wenn das Zeichen zum Beginn gegeben wird — auf der nächsten Seite ebenso mit den Buchstaben **b** und **j**. Streichen Sie dort also deutlich alle **b** und **j**. Arbeiten Sie dabei so sorgfältig und schnell wie möglich."

2. Probeprüfung

In Probeprüfungen, deren äußerer Verlauf schon im Zusammenhang mit dem vorigen Beispiel (S. 23) besprochen wurde, muß kontrolliert werden, ob die Instruktion verstanden wird.

Außerdem muß festgestellt werden, welche *Schwierigkeit* den Fähigkeiten der Probanden am besten angepaßt ist. Ist die Aufgabe zu einfach — z. B. beim Durchstreichen nur *eines* Buchstabens — so erfordert die Durchführung der Arbeit so geringe Anstrengung, daß der Ausfall vorwiegend tempobedingt ist. Ist sie dagegen zu schwer — wie etwa beim Durchstreichen von vier oder mehr Buchstaben oder von bestimmten Buchstabenverbindungen — so stellt die Lösung zu große Anforderungen an Gedächtnis und intellektuelle Fähigkeiten und die Arbeitsfaktoren als solche können nicht klar hervortreten.

Weiter muß untersucht werden, wieviel *Zeit* die einzelnen Probanden für die Bearbeitung des ganzen Testformulars benötigen, damit man später die Arbeitszeit richtig begrenzen kann. Man muß weiterhin klären, ob und wie stark die geforderte Leistung von Übungs- und Ermüdungsfaktoren beeinflußt wird, die sich bei den einzelnen Prüflingen erheblich unterscheiden und dementsprechend die Testergebnisse verschieden beeinflussen können.

Es würde hier zu weit führen, alle Variationen der Testbedingungen und der damit verbundenen Bewertungsmöglichkeiten, die in Vorversuchen geprüft werden, darzulegen. Ähnliche Fragen sind in früheren Forschungsarbeiten sehr eingehend experimentell untersucht worden, und man kann ihre Ergebnisse bei der Konstruktion und Anwendung des neubearbeiteten Tests verwerten. Diese Andeutungen sollen dem Leser nur zeigen, daß differenzierte und langdauernde Vorarbeiten notwendig sind, um zu einer brauchbaren Form des zunächst einfach erscheinenden Tests zu kommen.

3. Aufbau des Tests

Für den Aufbau des Durchstreichtests gibt es sehr verschiedene Möglichkeiten. Hier wird eine Fassung beschrieben, die auf Grund zahlreicher Probeprüfungen an Angestellten des mittleren Verwaltungsdienstes entwickelt worden ist. Sie enthält vier unmittelbar aufeinander folgende Teilaufgaben, die zunehmend schwieriger werden. In der ersten Aufgabe sind *zwei* Buchstaben (z. B. g und d) zu streichen, in der zweiten Aufgabe *drei* andere Buchstaben (z. B. h, y und f), in der dritten und vierten Aufgabe jeweils zwei Buchstaben mit Ausnahmen (z. B. alle *k*, außer wenn sie vor einem y stehen, und alle *q*, außer wenn sie nach einem d stehen). Die vierte Aufgabe unterscheidet sich noch dadurch von der dritten, daß während ihrer Bearbeitung ohne vorherige Ansage vom Prüfer irgendein interessanter Text verlesen wird.

Die Instruktion wird für alle vier Aufgaben *vor* der Bearbeitung des ganzen Tests schriftlich gegeben, damit die Prüflinge auf ein Zeichen hin (jeweils nach fünf Minuten) zur nächsten Aufgabe übergehen können. Es wird in der Instruktion darauf hingewiesen, daß man sich während des ganzen Tests durch nichts stören lassen soll.

Da mit einem in dieser Weise konstruierten Arbeitstest sehr verschiedene Leistungsfaktoren erfaßt werden sollen (s. S. 27), bedarf es einer sehr intensiven Analyse seiner Ergebnisse, bevor der Test für eine Eignungsprüfung herangezogen werden kann.

Denn es genügt nicht, daß man von einem theoretischen Ansatz — selbst wenn er noch so einsichtig erscheint — ausgeht und die Testergebnisse dann in dieser Richtung deutet. Man muß vielmehr für jeden gemessenen Leistungsfaktor die Frage der Zuverlässigkeit und Gültigkeit gesondert untersuchen. Wenn man z. B. die Umstellungsfähigkeit eines Prüflings nach der Art und Häufigkeit seiner Fehler zu Beginn jeder neuen Teil-Aufgabe bewerten will, so muß man diese Frage zunächst durch einen statistischen Vergleich mit entsprechenden Leistungen anderer Tests kontrollieren, bevor mit diesem Test gültige Aussagen über die Umstellungsfähigkeit gemacht werden können.

Neben dem Durchstreichtest kommen immer auch einige andere Arbeitstests zur Anwendung, so daß die Ergebnisse der verschiedenen Verfahren sich gegenseitig stützen und ergänzen können. Die Regeln für die Konstruktion dieser Verfahren sind die gleichen wie die an unseren beiden Beispielen dargestellten.

C.

Verfahren, die sich infolge von Konstruktionsmängeln für Gruppenprüfungen weniger eignen

Wenn der Aufbau eines Testverfahrens den Konstruktionsregeln nicht oder nur teilweise entspricht, ist eine zuverlässige Bewertung der Testleistung sehr in Frage gestellt. Unter den früher gebräuchlichen Verfahren sind einige, die solche Konstruktionsmängel aufweisen. Zum Beispiel gibt es Verfahren, die nur aus *wenigen* Einzelaufgaben bestehen. Obwohl sie zum Teil psychologisch sehr gut aufgebaut sind, eignen sie sich nur bedingt für eine quantitative Bewertung. Man kann sie mit einem Fachexamen vergleichen, bei dem das zu prüfende Gebiet nur von einer Seite her angegangen wird. Ein Kandidat kann in einem solchen Examen durch einen schlechten Ansatz oder durch Unsicherheit in dem betreffenden Einzelgebiet leicht seine ganze Examensleistung verderben, während ein anderer Glück hat und eine Frage vorgelegt bekommt, die ihm trotz geringer Kenntnisse zufällig geläufig ist. Niemand ist durch den Ausgang solcher Examina befriedigt.

Ähnliche Bedenken bestehen gegen die Verwendung mancher Organisationsaufgaben, Auftragstests usw., bei denen der Prüfling in einer gegebenen Zeit eine bestimmte Aufgabe lösen muß, deren Ertrag aber im Verhältnis zu der relativ langen Arbeitszeit gering ist. (Ein hoher Prozentsatz der Probanden liefert dabei — infolge eines ungeschickten Ansatzes — wenig brauchbare Ergebnisse.)

Ein Test soll also den geprüften Bereich von sehr verschiedenen Seiten und mit zahlreichen möglichst kurzen Aufgaben angehen. Wie groß ihre Zahl im Einzelfall sein muß, kann in den Probeprüfungen geklärt werden.

Es kommt jedoch nicht nur darauf an, daß ein Test *viele* Aufgaben enthält, sondern diese Aufgaben müssen — wenn sie quantitativ gewertet werden sollen — auch voneinander unabhängig sein. Bei der statistischen Analyse der Ergebnisse einer Reihe von Testverfahren hat es sich gezeigt, daß gerade die *Unabhängigkeit* der Einzelaufgaben ein sehr wichtiger Faktor für den Wert eines Tests ist.

In diesem Zusammenhang sind die sog. *Zuordnungstests* zu nennen, für die der *Sprichworttest* ein Beispiel ist. Der Prüfling soll dabei aus zwei Reihen von Sprichwörtern jeweils zwei sinnähnliche herausfinden und einander zuordnen. (Eine Fassung des Sprichworttests ist nachfolgend abgedruckt.)

a) Wo Honig ist, da finden sich die Bienen.
b) Ohne Fleiß kein Preis.
c) Es ist nicht alles Gold, was glänzt.
d) Ein alter Hund lernt nicht gern Kunststücke.
e) Auf kleine Streiche fällt selbst die Eiche.
f) Hochmut kommt vor dem Fall.
g) Wie man sich bettet, so liegt man.
h) Löwenmaul hat Hasenherz.
i) Lobe den Tag nicht vor dem Abend.
k) Blinder Eifer schadet nur.

1. Man soll ungelegte Eier nicht begackern.
2. Kleine Hunde bellen am meisten.
3. Wie man in den Wald ruft, so tönt es heraus.
4. Die Mühe muß vor dem Besitz kommen.
5. Erst wägen, dann wagen.
6. Biege die Weide, solange sie jung ist.
7. Glück gibt Gefährten.
8. Wer den Mund zu voll nimmt, muß viel schlucken.
9. Hohle Töpfe klappern am meisten.
10. Steter Tropfen höhlt den Stein.

(Lösung: a: 7, b: 4, c: 9, d: 6, e: 10, f: 8, g: 3, h: 2, i: 1, k: 5.)

Hier können — wie bei einem Silbenrätsel — Fehler oder Ungenauigkeiten am Anfang den weiteren Verlauf der Testlösung negativ beeinflussen, da zuletzt nur noch falsche Zuordnungen möglich sind und die Arbeit von neuem begonnen werden muß. Zu Sprichwort b) gehört z. B. Sprichwort Nr. 4; wählt ein Proband stattdessen etwa Nr. 10, so bleibt ihm später für Sprichwort e) kein Gegenstück, und die Verwirrung vergrößert sich.

Infolgedessen sollte man auch hier die zehn Aufgaben des Tests voneinander unabhängig gestalten, derart, daß für jedes Sprichwort der linken Seite vier bis fünf Sprichwörter zur Wahl gestellt werden, von denen der Prüfling das sinnähnlichste zu kennzeichnen hat.

Einige Fassungen des *Lückentests* gehören ebenfalls hierher, und zwar solche, bei denen die einzelnen Lücken sich — oft über mehrere Sätze hin — gegenseitig bedingen, also nicht unabhängig voneinander sind. Hierfür ein kurzer Ausschnitt aus einem schlechten Lückentest (für jede Textlücke soll ein Wort eingesetzt werden):

„Als sie auf hoher See ....., brach plötzlich ein starkes ..... aus, das erst ..... wurde, ... es schon die ..... der Mannschaft ..... hatte. Sofort ..... auf dem ganzen ..... Alarm gegeben. Die Mannschaftsräume waren jedoch ..... mehr zu ....."

Wenn ein Prüfling die zweite Lücke falsch löst, so müssen notwendig einige der folgenden Lücken ebenfalls falsch gelöst werden. Z. B.:

„Als sie auf hoher See *waren*, brach plötzlich ein starkes *Fieber* aus, das erst *bekämpft* wurde, als es schon die *Mehrzahl* der Mannschaft *befallen* hatte."

Es fällt dem Prüfling vielleicht erst beim Lesen des folgenden Satzes auf, daß er den Sinn nicht richtig aufgefaßt hat, und er kann nur unter erheblichem Zeitverlust den Fehler ausgleichen und mit der Lösung des Tests fortfahren. (Der Leser wird bei sich selbst beobachtet haben, wie schwierig es oft ist, von einem einmal gefaßten Denkansatz loszukommen und seinen Gedanken eine andere Richtung zu geben.)

Zum Vergleich die richtige Lösung dieser Textstelle:

„Als sie auf hoher See *waren*, brach plötzlich ein starkes *Feuer* aus, das erst *bemerkt* wurde, *als* es schon die *Räume* der Mannschaft *erfaßt* hatte. Sofort *wurde* auf dem ganzen *Schiff* Alarm gegeben. Die Mannschaftsräume waren jedoch *nicht* mehr zu *retten*."

Neben der Mehrdeutigkeit und der Abhängigkeit einzelner Lösungen voneinander haben verschiedene Lückentests den Mangel, daß sie in ihrem Stoff zu einseitig sind und dadurch bestimmte Wissens- und Interessengebiete begünstigen. Außerdem sind die in ihnen enthaltenen Lücken oft auf sehr verschiedene Weise lösbar (durch logisches Kombinieren, gedankliches Einfallsreichtum, sprachliche Gewandtheit, Kritikfähigkeit usw.), so daß die qualitative Analyse seiner Testleistung wesentlich ergiebiger ist als ein zahlenmäßiger Gesamtwert, dem der Gutachter nicht ansehen kann, auf welche Weise er entstanden ist.

Die heute angewandten Fassungen des Lückentests sind in der Regel statistisch so gut durchgearbeitet, daß die angeführten Fehlerquellen vermieden sind und die einzelnen Testleistungen zuverlässig bewertet werden können. Einige sind so aufgebaut, daß eine Anzahl voneinander unabhängiger Sätze aus den verschiedensten Sachgebieten zusammengestellt ist, in denen je ein oder zwei, in ihrer Schwierigkeit genau festgelegte und eindeutig lösbare Lücken enthalten sind.

Die Konstruktionsmängel eines Tests sind oft schon an der Verteilung seiner Ergebnisse zu erkennen. Im Zusammenhang mit der Testeichung („Häufigkeitsverteilung") wird darum nochmals auf einige der unzulänglich konstruierten Verfahren eingegangen werden.

D.

Arbeitszeit

Bei Gruppentests ist es unzweckmäßig, die Arbeitszeit jedes einzelnen Prüflings gesondert festzuhalten. Folgende Gründe sprechen dagegen:

1. Das Melden und Abgeben der Testformulare bringt eine starke Unruhe in den Ablauf der Prüfung. Die langsamer arbeitenden Prüflinge werden dadurch gestört, und die Güte ihrer Leistungen kann beeinträchtigt werden.
2. Schon bei einer Gruppe von 20 bis 30 Menschen ist es kaum möglich, die einzelnen Arbeitszeiten exakt festzuhalten, falls nicht mehrere Hilfskräfte zur Verfügung stehen. Bei noch größeren Gruppen leidet unvermeidlich die Objektivität der Durchführung unter der Zeitabnahme.
3. Die Arbeitszeiten differieren innerhalb einer Gruppe manchmal so stark, daß die schnellarbeitenden Prüflinge nach der Bearbeitung eines Tests längere Zeit untätig sind, ihre Leistungsmöglichkeiten also gar nicht ausnützen können. (Im Analogietest z. B. variierte die Arbeitszeit zwischen 2 und 20 Minuten.)
4. Einige Prüflinge scheuen sich, sofort nach Fertigstellung der letzten Aufgabe ihr Testformular abzugeben — und zwar aus den verschiedensten Gründen. Dadurch ergibt sich eine Häufung bei den langen Arbeitszeiten und eventuell eine ungenaue Bewertung der einzelnen Leistungen.

Durch das Festhalten der einzelnen Zeiten wird also nicht nur die *Durchführung* eines Tests, sondern auch die *Bewertung* seiner Resultate erschwert. In diesem Fall muß man die Arbeitszeiten in die Beurteilung einbeziehen, um den Leistungen der einzelnen Prüflinge gerecht zu werden.

Aus den angeführten Gründen ist man dazu übergegangen, bei vielen Gruppentests die Zeiten zu begrenzen, d. h. einen Test nach einer bestimmten genau festgelegten Zeitspanne abzubrechen und zum nächsten überzugehen. Auf diese Weise wird der Zeitfaktor konstant gehalten und der Ablauf der Prüfung einheitlich gestaltet.

Von Prüfungsteilnehmern oder Kritikern wird bisweilen der Einwand erhoben, daß solche Zeitbegrenzungen unberechtigt seien. Die langsamer arbeitenden Prüflinge — und das sind meistens die älteren — würden dabei benachteiligt, weil sie in der gegebenen Zeit weniger Aufgaben bearbeiten und damit auch richtig lösen als die, welche schneller und weniger gründlich vorgehen. Dieser Einwand ist unberechtigt, denn man kann einem Ergebnis durchaus ansehen, ob ein Prüfling langsam und gründlich oder etwa schnell und flüchtig gearbeitet hat. Außerdem hat man folgende Möglichkeiten, um die Leistung der einzelnen Prüflinge richtig zu beurteilen:

1. Man wird bei einer sehr uneinheitlichen Prüfgruppe, in der größere Alters- und Bildungsunterschiede bestehen, nie ausschließlich mit zeitbegrenzten Tests arbeiten, sondern immer auch einige freiere Verfahren einschalten, die Gelegenheit zum allmählichen Einarbeiten und gründlichen Durchdenken geben.
2. Um zu vermeiden, daß sich ein Prüfling bei einer Aufgabe zu lange aufhält und damit Zeit für das Lösen der folgenden Aufgaben verliert, wird *vor* Beginn einer schriftlichen Testprüfung darauf hingewiesen, daß in der jeweils gegebenen Zeit unmöglich der ganze Test bearbeitet werden kann, und daß die Probanden deshalb nur solche Aufgaben lösen sollen, die ihnen keine großen Schwierigkeiten bereiten.
3. In der *Einzelaussprache* ergibt sich die Möglichkeit, mit Prüflingen, die sich durch die Zeitbegrenzung benachteiligt fühlten, die einzelnen Testergebnisse zu besprechen.

4. Die Testleistungen werden nach besonderen *Altersnormen* bewertet; ein 47jähriger Prüfling wird beispielsweise nicht am Maßstab der gesamten Gruppe (oder gar der 20- bis 24jährigen), sondern an dem der 45- bis 49jährigen Berufsgruppe gemessen. (s. S. 50)

Die genaue Arbeitszeit für einen Test kann erst im Verlauf mehrerer, an verschiedenen Gruppen durchgeführter Probeprüfungen festgesetzt werden. — Sie darf nicht zu kurz sein, damit die einzelnen Resultate sich genügend voneinander abheben. Sie braucht andererseits nicht über ein bestimmtes Maß hinauszugehen, da die Verteilung der Ergebnisse sich dann — wie entsprechende Untersuchungen gezeigt haben — nur noch unwesentlich ändert, wenn nicht verschlechtert.

Tests, die aus Aufgaben von annähernd gleicher Schwierigkeit bestehen, müssen zu einem Zeitpunkt abgebrochen werden, an dem noch kein Proband sämtliche Aufgaben bearbeitet hat. Verlängert man bei diesen Tests die Arbeitszeit, so kommen zu viele Prüflinge zu hohen Werten und der Test differenziert unter den guten Leistungen nicht genügend. Hierfür spricht die Tatsache, daß bei vielen Fähigkeitstests (besonders bei Intelligenztests) für gute Leistungen in der Regel kürzere, für schlechte Leistungen längere Arbeitszeiten benötigt werden. — Im täglichen Leben ist es ähnlich: Der für eine bestimmte theoretische oder praktische Arbeit Begabte hat mit ihr weniger Schwierigkeiten und kommt im allgemeinen schneller zum Ziel als ein anderer, dem das betreffende Gebiet weniger liegt.

Ist ein Test in seiner Schwierigkeit abgestuft, so darf er erst dann abgebrochen werden, wenn etwa 90 Prozent der Probanden sämtliche Aufgaben bearbeitet haben. Denn hier müssen alle Prüflinge Gelegenheit haben, auch die schweren Aufgaben am Ende des Tests zu lösen, weil sonst die verschiedenen Ergebnisse nicht vergleichbar sind. Man setzt jedoch auch bei diesen Verfahren eine Zeitgrenze an, da ein Teil der Prüflinge erfahrungsgemäß nicht in der Lage ist, sämtliche Aufgaben zu lösen, und man sich mit ihnen bei der Prüfung unnötig aufhalten würde.

Die Arbeitsgeschwindigkeit als solche ist also bei einem gut konstruierten und geeichten Test für die Bewertung einer Testleistung nicht ausschlaggebend.

Zusammenfassend nochmals die wichtigsten *Regeln*, die bei der Konstruktion eines Tests beachtet werden müssen, wenn er eine quantitative Bewertung der Einzelleistungen ermöglichen soll:

1. Große Anzahl inhaltlich verschiedener Einzelaufgaben;
2. Unabhängigkeit der Aufgaben voneinander;
3. Eindeutigkeit und exakte Formulierung der Aufgaben (bei Wahllösungen am besten gewährleistet);
4. Gleiche oder gleichmäßig ansteigende Schwierigkeit der Aufgaben;
5. Unabhängigkeit von speziellen Kenntnissen und Interessen;
6. Anpassung in Stoff und Schwierigkeit an die zu prüfende Berufs-, Alters- und Bildungsschicht;
7. Klare, allgemeinverständliche Instruktion, die (wenn nötig mit Übungsbeispielen) schriftlich vorangestellt wird;
8. Durchführung der Probeprüfungen an repräsentativen Bevölkerungsausschnitten;
9. Eine dem Aufbau des Tests angepaßte Zeitbegrenzung;
10. Konstruktion von Paralleltests (soweit möglich).

Im vorliegenden Abschnitt wurde versucht, dem Leser deutlich zu machen, daß die Konstruktion eines Tests eine recht umständliche und zeitraubende Angelegenheit ist, deren Gelingen nicht zuletzt von dem Entgegenkommen derjenigen Kreise, die berufsmäßig mit Personalfragen zu tun haben, sowie von der Bereitwilligkeit der Probanden abhängt. Denn es ist nicht immer leicht, die für die Konstruktion eines Tests notwendige Anzahl und Auswahl von Probanden zu finden und genügend Zeit für wiederholte Probeprüfungen zur Verfügung zu haben.

Die Konstruktion ist aber nur die erste Stufe bei der Entwicklung eines neuen Verfahrens. Denn selbst ein gut aufgebauter Test ist praktisch wertlos, so lange man nicht weiß, nach welchen Richtlinien man ihn auswerten soll und welche Bedeutung die Leistungen der einzelnen Prüflinge haben. Der Klärung dieser Fragen dient die sogenannte Eichung eines Tests.

IV.

Eichung von Gruppentests

Das Ziel der Eichung sind Normen, die die objektive Beurteilung der Einzelleistung ermöglichen. Dazu bewertet man die Testleistungen einer großen und repräsentativen Anzahl von Probanden, ordnet die Resultate (Werte) und bearbeitet sie nach statistischen Gesichtspunkten weiter. Diese Schritte sollen jetzt durch Beispiele erläutert werden.

A. Bewertung (Problem der Objektivität)

Wenn hier über die Bewertung von Gruppentests gesprochen wird, so ist damit — dem Rahmen dieser Arbeit entsprechend — lediglich die *quantitative* Bewertung gemeint, die zur Feststellung eines zahlenmäßigen Ergebnisses für die Testleistung jedes Probanden führt. Wie dieser Wert im Einzelfall zustande kommt und was er im Leistungsbild einer Persönlichkeit bedeutet, wird darin nicht ausgedrückt.

Damit die Werte so zuverlässig und treffend wie möglich sind, muß die Bewertung der Testleistungen möglichst *objektiv*, d. h. frei von willkürlichen Entscheidungen des Auswerters erfolgen. Außerdem muß eine geeignete, dem Test angepaßte *Bewertungsmethode* angewandt werden.

1. Subjektive und objektive Bewertung

Es hat sich erwiesen, daß die Bewertung eines Tests um so zuverlässiger ist, je eindeutiger und enger begrenzt die geforderten Lösungen sind. Beispiele aus verschiedenen Testverfahren sollen das verdeutlichen:

I) Im *Zahlenreihentest* sind Reihen von Zahlenfolgen gegeben, die nach einer bestimmten Regel aufgebaut sind. Der Prüfling soll diese Regeln erkennen und die Reihen dementsprechend jeweils um zwei Zahlen fortsetzen. Z. B.:

									Lösung:
a)	9	7	8	6	7	5	.....	.....	*6 4*
b)	4	5	8	9	12	13	.....	.....	*16 17*

II) Die gleiche Aufgabe soll — mit anderem Material — im *Buchstabenreihentest* gelöst werden. Z. B.:

									Lösung:
a)	b	d	f	h	j	l	.....	.....	n p
b)	u	r	p	m	k	h	.....	.....	f c

III) Der *Denkrechentest* enthält eingekleidete Aufgaben, in denen Zahlenbeziehungen gefunden werden sollen. Z. B.:

a) Wenn man unter 2 Personen (A und B) 130 DM so verteilt, daß A 40 DM mehr erhält als B, wieviel bekommt dann A? *(85,— DM)*
b) Wieviel Würfel von 1 ccm Größe kann man in eine Schachtel tun, die 5 cm breit, 7 cm lang und 2 cm hoch ist? *(70 Stück)*

Für die Aufgaben dieser Tests gibt es nur *eine* richtige Lösung. Der Auswerter hat also lediglich zu kontrollieren, ob die Ergebnisse richtig oder falsch sind; ihm fällt keinerlei Entscheidung über den Wert einer Lösung zu. Die in dieser Art konstruierten Tests werden darum als *objektive* Tests bezeichnet. Man kann sie — wie bei der Konstruktion des Analogietests gezeigt worden ist (s. S. 26) — auch nach dem Prinzip der Wahllösungen aufbauen, indem man für jede Aufgabe mehrere Lösungen zur Auswahl gibt. Das ist in jedem Fall bei solchen Tests erforderlich, deren Aufgaben mehrere richtige Lösungen zulassen. Auch hierfür einige Beispiele:

IV) Im *Fremdworttest* soll der Prüfling für eine Reihe von Fremdwörtern die deutsche Übersetzung finden. Dazu sind für jedes Wort fünf Lösungen zur Wahl gestellt, unter denen die sinnentsprechende gekennzeichnet werden soll. (Dieser Test prüft also vorwiegend das Bildungsniveau.) Z. B.:

a) POSITION	1. Vorbild	c) KONSUM	1. Verbrauch
	2. Umstand		2. Umsatz
	3. Ergebnis		3. Bestand
	4. Stellung		4. Vorrat
	5. Dienstgrad		5. Bedarf
b) REVISION	1. militärische Einheit	d) KONFLIKT	1. Zusammenkunft
	2. Nachprüfung		2. Widerspruch
	3. Abänderung		3. Streitgespräch
	4. Wiedersehen		4. Zwiespalt
	5. Berufung		5. Aufforderung

(Lösungen: a: 4, b: 2, c: 1, d: 4.)

V) Bei der *Textanalyse* muß aus fünf zur Wahl gestellten Sätzen (a—e) derjenige herausgefunden werden, der seinem Sinn nach in einem gegebenen Text enthalten ist. Z. B.:

„Als sich das Anwachsen der Bevölkerung ernstlich bemerkbar machte und die Kosten für die Rohstoffe zu steigen begannen, richtete man die Aufmerksamkeit auf die Erhaltung der noch vorhandenen Rohstoffquellen. Untersuchungen zeigten, daß die Vorräte an Holz und Mineralien nicht unerschöpflich sind, sondern daß im Gegenteil ihre Menge begrenzt ist."

a) Die Wichtigkeit der Erhaltung von Rohstoffquellen ist erst in den letzten Jahren erkannt worden.
b) Die Lebensdauer der aus Holz und Metall hergestellten Gegenstände ist begrenzt.
c) Der Rohstoffmangel hat das Anwachsen der Bevölkerung beeinträchtigt.
d) Die Begrenztheit unserer Rohstoffquellen hat zu Maßnahmen geführt, durch die unsere Rohstoffe geschont werden sollen.
e) Unsere Holzvorräte sind nahezu erschöpft.

(Hier ist die Lösung d), da dieser Satz sinngemäß dem darüber stehenden Text entspricht.)

Bei den in dieser Form aufgebauten Tests ist die Objektivität der Bewertung gesichert, da für eine subjektive Beurteilung der Lösungen keinerlei Raum gelassen ist. Auch die technische Seite der Auswertung ist dabei vereinfacht, die Schwierigkeit des Entzifferns handgeschriebener Lösungen fällt fort.

Den objektiven Tests, wie sie in den Beispielen I bis V veranschaulicht wurden, kann man die *subjektiven* Verfahren gegenüberstellen. Da ihre Aufgaben sehr verschiedene Arten der Lösung zulassen, muß es bis zu einem gewissen Grade der Beurteilung des Auswerters überlassen bleiben, wie hoch er den Wert einer Lösung einschätzt. Einige Beispiele für subjektive Verfahren:

VI) Um den Gegensatz zu den oben angeführten Testbeispielen besonders deutlich zu machen, sei zunächst ein *Telegrammtest* wiedergegeben, bei dem die Subjektivität infolge von Konstruktionsmängeln noch vergrößert wird, und der deshalb für eine zuverlässige quantitative Bewertung ungeeignet ist. An diesem Beispiel wird der Leser erkennen, wie stark und für den Wert des Verfahrens abträglich der Einfluß ist, der hier der subjektiven Beurteilung durch den Auswerter eingeräumt ist.

Instruktion:

Die nachstehenden Mitteilungen sind in einem Telegramm zusammenzufassen! Alles Wichtige muß darin festgehalten werden! Das Telegramm darf nicht mehr als 25 Wörter enthalten. Sie haben dafür 8 Minuten Zeit.

Text des Briefes:

Lieber Egon! Kommenden Freitag wollte ich verabredungsgemäß mit unseren Ausschußmitgliedern in Stuttgart die Arbeiten der nächsten Wochen besprechen. Nun mußte ich aber hier in Ulm auf diesen Freitagabend eine wichtige Versammlung ansetzen. Die vorgesehene Besprechung muß daher vertagt werden. Ich komme nun eine Woche später wie verabredet nach Stuttgart. Verständige bitte die Ausschußmitglieder. Ich kann mich doch darauf verlassen? Ich bitte Dich, doch der Versammlung in Ulm auf alle Fälle beizuwohnen. Es ist wirklich dringend. Wenn Du 14.25 Uhr in Stuttgart abfährst, kannst Du in Göppingen Herrn Schuster mitnehmen, der mit mir bereits telefonierte und versprach zu kommen. Ich erwarte Euch beide im Parteilokal um 16.00 Uhr. Die Versammlung ist so wichtig, daß Du unbedingt teilnehmen mußt.

Auf Wiedersehen am Freitag und inzwischen herzliche Grüße von

Deinem E r i c h.

Es kommt natürlich zu den verschiedensten Lösungen, von denen hier nur einige herausgegriffen seien.

Lösungen:

a) Verschiebe Besprechung Stuttgart um eine Woche, verständige Ausschußmitglieder. Erwarte Dich und Schuster Freitag 16.00 Uhr in Ulm Parteilokal in wichtiger Angelegenheit.

b) Besprechung Stuttgart nicht möglich. Dein Beiwohnen erforderlich. Verständige Ausschuß. Ankomme Stuttgart wie verabredet. — Schuster nach Ulm mitbringen — Erwarte Euch Parteilokal — Deine Teilnahme erforderlich. Erich.

c) Muß Besprechung Ausschußmitglieder verschieben um eine Woche. Verständige Mitglieder. Erwarte Dich Freitag Ulm ab Stuttgart 14.25 in Göppingen Schuster mitnehmen. Erwarte Euch 16.00 Parteilokal.

d) Ausschußmitgliederbesprechung 1 Woche später. Ausschußmitglieder verständigen abfährst Stuttgart 14.25 Schuster aus Göppingen mitbringen erwarte Euch Parteilokal 16.00.

e) Versammlung vertagen. Besprechung Freitag in Ulm. Teilnahme mit Schuster erforderlich. Abreise 14.25 über Göppingen. Dort Schuster mitnehmen. Wiedersehen Parteilokal 16.00.

f) Vorgesehene Besprechung auf Freitag nächster Woche vertagt, Ausschußmitglieder verständigen. Diesen Freitag Versammlung Ulm beiwohnen. Abfahrt Stuttgart 14.25, Schuster Göppingen steigt zu. Eintreffen Parteilokal 16.00. Erich.

g) Besprechung am kommenden Freitag. Bitte ab Stuttgart 14.25. Herrn Schuster Göppingen mitbringen. Ankunft Parteilokal 16.00. Vorgesehene Besprechung vertagt, veranlasse das Nötige.

usw.

Man stelle sich den Auswerter vor, der in möglichst kurzer Zeit 40 bis 50 solcher Lösungen beurteilen soll. Selbst wenn genaue Bewertungsrichtlinien ausgearbeitet sind — wenn z. B. die im Brief enthaltenen wesentlichen Punkte herausgelöst und mit verschiedenen Gewichten belegt werden —, können Ungenauigkeiten oder Irrtümer bei der Bewertung nie ganz vermieden werden. Denn die einzelnen Punkte bedingen sich z. T. gegenseitig, außerdem kann die Güte einer Lösung durch sprachliche Gewandtheit, durch Flüssigkeit oder Unbeholfenheit des Stils wesentlich beeinflußt werden. Einem zahlenmäßigen Resultat kann man kaum ansehen, ob es vorwiegend durch logisches Denken, durch Sprachgewandtheit oder durch Routine im Aufsetzen von Telegrammen zustande gekommen ist. Mit diesen Einwänden wird die Brauchbarkeit eines solchen Tests in Einzel- oder kleineren Gruppenprüfungen, bei denen die qualitative Beurteilung der individuellen Denkform und des Stils im Vordergrund stehen, nicht schlechthin verneint; von dem zahlenmäßigen Ergebnis allein darf die Beurteilung der Testleistung jedoch nicht abhängen.

Für den Leser, den eine quantitative Bewertung der oben abgedruckten Lösungen interessiert und der vielleicht versucht hat, den Wert dieser Lösungen selbst einzuschätzen, sei eine Rangordnung angeführt, die ihnen durch eine Gruppe von Auswertern gegeben wurde. (Die Rangplätze wurden vom Gesamteindruck her geschätzt):

Lösungen:	f	c	e	a	d	g	b
Rangplätze:	1.	2.	3.	4.	5.	6.	7.

Die Lösung f wurde also als beste, die Lösung b als schlechteste Testleistung gewertet. Der Abstand zwischen den einzelnen Rangplätzen (1—7) war dabei nicht gleich groß, die drei mittleren Lösungen lagen näher zusammen als die extremeren.

Man hat auch den Versuch einer Punktbewertung nach bestimmten Richtlinien gemacht mit folgendem Ergebnis:

Lösungen:	f	c	a	e	d	g	b
Punktzahl:	19	14	14	10	7	7	0

Der Vergleich beider Bewertungen zeigt, daß zwar die beste (f) und die schlechteste Lösung (b) klar erkannt wurden (wozu es übrigens kaum einer besonderen Bewertungsmethode bedarf), daß aber die Bewertung der hier wesentlich interessierenden Mittellösungen nur mit Vorbehalt für die Beurteilung herangezogen werden kann.

Es gibt freilich eine Reihe anderer subjektiver Tests, bei denen die im vorigen Beispiel geschilderten Konstruktionsmängel (vor allem die gegenseitige Abhängigkeit der einzelnen Aufgaben) fortfallen, deren Bewertung jedoch wegen der sehr verschiedenen Lösungsmöglichkeiten ebenfalls nie ganz objektiv erfolgen kann. Hierher gehören Definitionstests wie z. B. die subjektive Fassung des Fremdworttests.

VII) Im *Fremdwort-Definitionstest* wird die begriffliche Bestimmung von zehn geläufigen Fremdwörtern verlangt, die in ihrer Schwierigkeit dem Bildungsniveau der Prüflinge angepaßt sind. Einige Beispiele für solche Begriffsbestimmungen, wie sie von einem Probanden gegeben wurden:

Konflikt ist ein Zwiespalt, der sowohl innerer Natur sein kann als auch äußerlich zwischen zwei Parteien bestehen kann.

Interesse ist die gesteigerte Anteilnahme an einer Sache oder Person.

Opposition bedeutet Widerstand — hauptsächlich in geistiger oder politischer Hinsicht.

Liquidation ist die Auflösung und finanzielle Abwicklung einer Firma, Gesellschaft.

usw.

Der Vergleich dieser Beispiele, die aus einem an Verwaltungsangestellten gewonnenen Testmaterial entnommen sind, mit dem unter IV angeführten Fremdworttest läßt deutlich den Unterschied zwischen subjektiven und objektiven Tests erkennen. Bei den *objektiven* Tests liegt für die Prüfstelle die Hauptarbeit in der Konstruktion des Verfahrens, also *vor* seiner Anwendung. Bei den *subjektiven* Tests dagegen besteht die Hauptarbeit in der Bewertung der einzelnen Lösungen, sie erfolgt also *nach* der Anwendung des Tests. Um die Auswertung dieser Tests, bei denen dem Prüfling die Formulierung der Lösungen mehr oder weniger freigestellt ist, so weit wie möglich zu objektivieren, müssen oft sehr differenzierte Bewertungsmethoden angewandt werden. Trotzdem läßt sich die durch den subjektiven Faktor mögliche Fehlerquelle nie völlig ausschalten. Aus diesem Grunde werden die früher fast ausschließlich gebrauchten subjektiven Tests in schriftlichen Gruppenprüfungen jetzt weitgehend durch objektive Tests ersetzt.

In bestimmten Fällen wird man allerdings auf subjektive Verfahren nicht ganz verzichten können, vor allem dann, wenn man die Selbständigkeit und persönliche Eigenart des Denkens erfassen will. Das gilt besonders für die Prüfung von Bewerbern für selbständige leitende Positionen, für wissenschaftliche Tätigkeiten (z. B. Studienanwärter) oder für Berufe, in denen

sprachliche Ausdrucksgewandtheit gefordert wird (z. B. selbständige Sekretärinnen). In diesen Fällen muß die Objektivität auf andere Weise angestrebt werden. Es ist dabei üblich, daß mehrere geschulte Beurteiler die Testunterlagen unabhängig voneinander nach genauen Bewertungsregeln auswerten. (Die Zahl dieser Beurteiler hängt davon ab, wie stark die verschiedenen Bewertungen bei einem Test differiert haben.) Den einzelnen Testleistungen wird derjenige Wert zugesprochen, der sich aus einer statistischen Kombination der verschiedenen Urteile ergibt.

Im weiteren Verlauf dieser Arbeit wird die Eichung *objektiver* Tests dargestellt, während die subjektiven Verfahren nur gelegentlich zum Vergleich herangezogen werden.

2. Gegenstand der Bewertung und Bewertungsmethoden

Nachdem das Problem der objektiven Bewertung an verschiedenen Tests gezeigt worden ist, muß jetzt auf die Frage eingegangen werden, *was* bei einem Test bewertet wird, d. h. wie man die Werte ermittelt, nach denen die einzelnen Testleistungen beurteilt werden. Hierfür bieten sich — je nach der Art des Tests — vier Möglichkeiten:

1. Die Werte entsprechen der *Menge* der in einer bestimmten Zeit geleisteten Arbeit.
2. Sie entsprechen der *Zeit*, die für eine bestimmte Arbeit benötigt wurde.
3. Sie richten sich nach der *Güte* der einzelnen Lösungen.
4. Sie richten sich nach der *Schwierigkeit* der richtig gelösten Einzelaufgaben.

Zu 1) Die erste Möglichkeit, daß nämlich die *Menge der geleisteten Arbeit* gewertet wird, ergibt sich, wenn die Arbeit durchgehend ungefähr gleich schwer ist, wenn z. B. ein Test aus Einzelaufgaben von annähernd gleicher Schwierigkeit zusammengesetzt ist. Als „geleistete Arbeit" gilt dabei die Summe der fehlerfrei gelösten Aufgaben oder der richtig bearbeiteten Elemente.

Ist also ein Analogietest aus 30 Aufgaben von annähernd gleicher Schwierigkeit aufgebaut, so wird das Resultat durch die Summe der in einer bestimmten Zeit fehlerfrei gelösten Aufgaben ausgedrückt. Das Resultat eines Probanden kann demnach 0—30 Lösungen betragen.

Im Durchstreichtest werden die Werte durch die Anzahl der in einer festgesetzten Zeit (5 Minuten je Aufgabe) richtig durchstrichenen Buchstaben bestimmt.

Zu 2) Die zweite Bewertungsmöglichkeit, welche die *Arbeitszeit* für den Wert einer Testleistung entscheiden läßt, ist bei Gruppenprüfungen wenig geeignet, weil sie das Notieren der einzelnen Arbeitszeiten erforderlich macht.

Bei kleineren Gruppenprüfungen oder Einzeluntersuchungen ist es manchmal noch üblich, jeden Test bis zum Ende bearbeiten zu lassen und die Arbeitszeiten einzeln festzuhalten. In diesem Falle müssen Güte *und* Menge gewertet werden, denn der Wert einer Leistung steigt selbstverständlich mit der Kürze der dafür benötigten Zeit. Die beiden Werte sollten jedoch nicht in einer Zahl zusammengefaßt werden, da die Eigenart der Leistung in einem solchen Gesamtwert nicht richtig zum Ausdruck kommen kann. — Güte und Arbeitszeit sind verschiedene und nicht die einzigen Seiten einer Leistung. Für den Gutachter ist der Vergleich beider Aspekte wertvoller als eine undifferenzierte Gesamtzahl. Der Statistiker kann in manchen Fällen bestimmte Techniken anwenden, um die verschiedenen Werte zu kombinieren.

Zu 3) Eine Bewertung nach der *Güte der einzelnen Leistungen* setzt voraus, daß verschieden gute Lösungen möglich sind. Sie kommt also im wesentlichen bei subjektiven Tests in Frage.

Wenn z. B. bei einem Definitionstest die einzelnen Definitionen nur danach bewertet werden, ob sie richtig oder falsch sind, wird man den Leistungen der Probanden nicht gerecht, denn die Lösungen können inhaltlich und formal sehr verschieden gut sein. Außerdem ist die Grenze zwischen eindeutig richtigen und falschen Lösungen nicht scharf zu ziehen. — In diesem Fall muß man den einzelnen Lösungen also, je nachdem wie umfassend und präzise sie formuliert sind (begriffliche Klarheit, Ausdrucksvermögen), verschiedene Werte geben. — Um die Entscheidung

des Auswerters bei dieser differenzierten Bewertung möglichst einzuschränken, werden vor der Anwendung eines solchen Tests durch mehrere geschulte Beurteiler an Hand des Probematerials *Bewertungsskalen* angelegt. Darin sind für jeden zu definierenden Begriff verschiedene charakteristische Lösungsbeispiele zusammengestellt, die nach Güteklassen geordnet sind. Die einzelnen Güteklassen (etwa vier oder fünf) werden mit verschiedenen Punktwerten belegt.

Durch den Vergleich mit diesen Skalen können die Testleistungen bewertet werden. Dabei arbeiten mehrere Auswerter getrennt voneinander und ihre Ergebnisse werden anschließend miteinander verglichen.

Auf diese Weise lassen sich für die subjektiven Abweichungen der Bewertung enge Grenzen setzen, ohne daß die Eigenart des Tests dabei verliert.

Zu 4) Die Bewertung nach der *Schwierigkeit der Aufgaben* schließlich ist bei solchen Tests möglich, deren Einzelaufgaben in ihrer Schwierigkeit abgestuft sind (s. S. 25). Man kann dann ein entsprechend gestaffeltes *Punktsystem* ausarbeiten, den richtig gelösten Aufgaben also verschiedene Punktwerte geben. Diese Methode ist jedoch nur dann zulässig, wenn ein Test sehr exakt konstruiert ist; vor allem müssen die Schwierigkeitsunterschiede der Aufgaben statistisch gesichert sein.

Im Verlauf der Entwicklung eines Tests, die sich gewöhnlich über Jahre hinzieht, geht man bei der Bewertung oft von verschiedenen Ansätzen aus, bis man die zuverlässigste und zugleich technisch einfachste Methode gefunden hat. — So hat man früher gelegentlich die Testleistungen nach der Zahl oder dem Gewicht der *Fehler* bewertet. Bei Gruppentests ist das schon deshalb nicht möglich, weil die einzelnen Verfahren nicht bis zum Ende bearbeitet werden, man also an der Fehlerzahl nicht die wirkliche Leistung erkennen kann. Außerdem ist es in jedem Fall sinnvoller, eine Leistung nicht negativ (durch die Anzahl der Fehllösungen), sondern positiv auszudrücken. Auch kompliziertere Methoden, wie etwa die Berechnung eines „Fehlerprozents" (die Zusammenfassung von richtigen und falschen Lösungen in einem Wert) haben sich fast immer als unergiebig erwiesen.

Bei der Entwicklung des Durchstreichtests z. B. wurden verschiedene Bewertungsmethoden erprobt. — Der Auswerter stellt für jede Teilaufgabe die Zahl der Fehler, also der beim Durchstreichen übersehenen Buchstaben fest und notiert daneben die Anzahl der bearbeiteten Zeilen. Würde man der Bewertung die Fehlerzahl zugrunde legen, so wären die einzelnen Leistungen nicht richtig beurteilt, da die Probanden in der gegebenen Zeit sehr verschieden viele Buchstaben bearbeiten, also auch verschieden viele Fehlermöglichkeiten haben (5 Fehler haben in 10 Zeilen natürlich ein anderes Gewicht als etwa in 30 Zeilen). Die Fehler kann man nur dann absolut vergleichen, wenn jeder Proband die gleiche Menge, z. B. 30 Zeilen, bearbeitet hat.

Um diese Schwierigkeit auszuschalten, stellte man für jede Einzelleistung ein Fehlerprozent fest, d. h. man berechnete den prozentualen Anteil der Auslassungen an der Gesamtzahl der Buchstaben, die in dem bearbeiteten Abschnitt zu streichen waren. Ganz abgesehen davon, daß diese Berechnung (jeweils für vier Teilaufgaben gesondert) ziemlich umständlich ist, ergab sich auch eine ungünstige Verteilung der so erhaltenen Werte, da die niedrigen Prozentwerte viel häufiger vorkommen als die höheren (s. S. 45 Abb. 12).

Schließlich wurde die Zahl der in der festgelegten Zeit richtig gestrichenen Buchstaben, also die positiv geleistete Arbeit, bewertet. Hierbei ergab sich eine überraschend gute Verteilung der einzelnen Werte, und statistische Kontrollmethoden zeigten, daß die Testleistung mit diesen Werten am zuverlässigsten erfaßt wird[1]). — Auf die Sorgfalt oder Flüchtigkeit der Arbeit können aus dem Vergleich von richtig gestrichenen und insgesamt bearbeiteten Buchstaben Schlüsse gezogen werden.

Es hat sich gezeigt, daß bei objektiven Fähigkeitstests, wie sie in größeren Gruppenprüfungen vor allem zur Anwendung kommen, in der Regel die einfachste Bewertungsmethode, nämlich das Zählen der richtigen Lösungen oder der richtig bearbeiteten Elemente, die beste ist, während bei subjektiven Tests meist ein differenzierteres Bewertungsschema ausgearbeitet werden muß.

[1]) vgl. Meili a. a. O. S. 69

Daß in vielen Fällen, besonders bei Arbeitstests, das zahlenmäßige Ergebnis für die *psychologische* Beurteilung einer Testleistung nicht ausreicht, ist schon erwähnt worden. Ein Testwert drückt immer nur die *positive* Leistung aus. Erfolglose Arbeit — wie Fehllösungen oder Verzögerungen durch Nichtfinden der Lösung — wird nicht in das zahlenmäßige Ergebnis selbst einbezogen. So muß die Zahl der Fehler und der nicht gelösten Aufgaben getrennt angegeben werden, damit der Gutachter die Möglichkeit hat, bestimmte Rückschlüsse auf die Arbeitshaltung eines Prüflings (Sorgfalt, Großzügigkeit usw.) zu ziehen.

Diese Problematik wird an folgendem Beispiel deutlich:
Wenn im Analogietest ein Prüfling in der gegebenen Zeit 15 richtige und keine falschen Lösungen hat, so kann das für den Gutachter etwas anderes bedeuten, als wenn ein zweiter 20 Aufgaben richtig und 5 falsch gelöst hat. — Wollte man argumentieren, die erste Leistung sei besser, weil bei ihr 100 Prozent richtige Lösungen vorliegen gegenüber nur 80 Prozent richtigen Lösungen des zweiten Prüflings, so würde dadurch das Verhältnis dieser beiden Leistungen in unzulässiger Weise verschoben.

Zusammenfassend gilt für die quantitative Bewertung von Tests:

1. Die Bewertung ist um so einfacher, sicherer und zuverlässiger, je gründlicher ein Test vorher durchgearbeitet worden ist.
2. Objektive Tests, besonders wenn sie nach dem Prinzip der Wahllösungen aufgebaut sind, ermöglichen die zuverlässigste Bewertung.
3. Subjektive Tests erfordern eine Objektivierung der Bewertung durch genaue Bewertungsregeln (Güteklassen-Skalen) und durch den Einsatz mehrerer Auswerter.
4. Man soll tunlich die *positiven* Leistungen bewerten und nicht von den Fehlern oder von einer Kombination verschiedener Leistungsgesichtspunkte ausgehen.

B.
Aufbereitung des Materials

Für die Eichung eines Tests muß zunächst festgestellt werden, wie sich die Ergebnisse einer umfangreichen und repräsentativen Gruppe von Probanden auf die verschiedenen Leistungsstufen verteilen. Diese „Häufigkeitsverteilung" des Tests muß bekannt sein, weil sich erst dann bestimmen läßt, ob die gebräuchlichen technischen Methoden bei der Eichung anwendbar sind.

1. Häufigkeitsverteilung

Nach der Auswertung werden die Testergebnisse der einzelnen Probanden in der Reihenfolge der Prüfnummern in die *Urliste* eingetragen. (Eine Urliste wird für jeden Test gesondert angelegt).

Beispiel: Urliste Durchstreichtest

Lfd.	Teilaufgabe I			II			III			IV		
Nr.	Z	R	F	Z	R	F	Z	R	F	Z	R	F
1	18,7	138	6	10,5	108	11	6,2	38	8	7,2	47	6
2	21	137	26	17,2	175	21	6,4	42	5	8	53	8
3	21,2	163	2	14,4	149	10	6	37	7	7,9	45	14
4	23,5	176	7	16,7	173	17	9,2	62	5	9	43	28
5	28,9	203	20	15,3	152	19	8	49	10	5,10	31	13

usw.

Z = Zeilenzahl
R = richtige Durchstreichungen
F = Fehler (übersehene Buchstaben und falsche Durchstreichungen)
Der Ausschnitt aus der Urliste für den Durchstreichtest zeigt die Testwerte einiger Probanden (1—5) in den 4 Teilaufgaben. Es sind jeweils die Zeilenzahl, die Anzahl der richtigen Durchstreichungen und die der Fehler eingetragen. Für die Eichung müssen die Werte von mindestens 200 Probanden der betreffenden Bevölkerungsschicht vorliegen.

Um mit einer großen Zahl von Werten arbeiten zu können, muß man diese zunächst systematisch ordnen. Dazu wird eine *„Häufigkeitstabelle"* angelegt, die zeigen soll, wie oft jeder Wert vorgekommen ist. Sie wird begrenzt durch den höchsten und den niedrigsten der in der Urliste enthaltenen Werte. Läßt ein Test viele verschiedene Wertmöglichkeiten zu, so werden immer mehrere aufeinanderfolgende Werte zu einer *Klasse* zusammengefaßt, so daß die Skala der Häufigkeitstabelle insgesamt etwa 10—20 Klassen enthält.[1]) In die Häufigkeitstabelle werden die einzelnen Werte der Urliste als Striche eingetragen. Die Zahl der Resultate, die in eine Klasse fallen, wird als „Klassenhäufigkeit" (f) bezeichnet. Die Summe aller Klassenhäufigkeiten entspricht der Gesamtzahl der Probanden (N), die den Test bearbeitet haben. Die auf diese Weise gewonnenen Daten bilden die *Häufigkeitsverteilung* des betreffenden Tests.

Klassengrenzen	Klassenhäufigkeiten	f	f %
260–269	I	1	0,25
250–259	II	2	0,5
240–249	IIII	4	1,0
230–239	IIII IIII	9	2,25
220–229	IIII II	7	1,75
210–219	IIII IIII III	13	3,25
200–209	IIII IIII IIII I	21	5,2
190–199	IIII IIII IIII IIII III	28	6,9
180–189	IIII IIII IIII IIII I	31	7,7
170–179	IIII IIII IIII IIII IIII IIII I	41	10,1
160–169	IIII IIII IIII IIII IIII IIII III	39	9,7
150–159	IIII IIII IIII IIII IIII IIII I	41	10,1
140–149	IIII IIII IIII IIII IIII IIII III	39	9,7
130–139	IIII IIII IIII IIII IIII IIII IIII	40	9,9
120–129	IIII IIII IIII IIII IIII IIII	35	8,7
110–119	IIII IIII IIII IIII	20	5,0
100–109	IIII IIII III	14	3,5
90 – 99	IIII III	10	2,5
80 – 89	IIII	5	1,25
70 – 79	II	2	0,5
60 – 69	I	1	0,25
	N =	403	100,00 %

Abb. 1 Häufigkeitstabelle (Strichliste) Durchstreichtest I

X	Häufigkeiten	f	f %
20	IIII I	6	1,2
19	IIII IIII IIII	14	2,9
18	IIII IIII IIII IIII	20	4,1
17	IIII IIII IIII IIII II	27	5,6
16	IIII IIII IIII IIII IIII	29	6,0
15	IIII IIII IIII IIII IIII III	33	6,8
14	IIII IIII IIII IIII	25	5,2
13	IIII IIII IIII IIII IIII III	34	7,0
12	IIII IIII IIII IIII III	33	6,8
11	IIII IIII IIII IIII I	26	5,4
10	IIII IIII IIII IIII III	28	5,8
9	IIII IIII IIII IIII IIII I	36	7,4
8	IIII IIII IIII IIII	25	5,2
7	IIII IIII IIII IIII IIII	29	6,0
6	IIII IIII IIII IIII IIII II	32	6,6
5	IIII IIII IIII III	23	4,7
4	IIII IIII IIII II	23	4,7
3	IIII IIII IIII II	22	4,5
2	IIII IIII II	12	2,5
1	IIII	5	1,0
0	III	3	0,6
	N =	485	100,0 %

Abb. 2 Häufigkeitstabelle (Strichliste) Analogietest

Zu Abbildung 1:

Im Durchstreichtest (Teilaufgabe I) lagen die beiden Extremwerte der Urliste bei 62 und 267 richtigen Durchstreichungen. Man muß deshalb mehrere Werte (etwa 10, 15 oder 20) zu einer Klasse zusammenfassen. In der hier abgebildeten Häufigkeitstabelle enthält jede Klasse 10 Werte. Das Resultat jedes einzelnen Probanden wird als Strich in die betreffende Klasse eingetragen (der fünfte Strich jeweils quer durch die vier vorhergehenden). Die einzelnen Klassenhäufigkeiten sind unter f in der Tabelle abzulesen.

Zu Abbildung 2:

Die hier als Beispiel gewählte Fassung des Analogietests enthält 20 Aufgaben, also 21 Wertmöglichkeiten (0—20 richtige Lösungen). Die Eintragung der einzelnen Resultate erfolgt wie im ersten Beispiel.

Zur anschaulichen Beurteilung einer Häufigkeitsverteilung können die Daten der Häufigkeitstabelle in eine *graphische Darstellung* übertragen werden.

Die Voraussetzungen für das Verständnis graphischer Darstellungen — soweit man sie zur Veranschaulichung von Testleistungen heranzieht — sollen kurz in Erinnerung gebracht werden:

Um eine graphische Darstellung anzulegen, bedient man sich eines Koordinatensystems, das aus zwei rechtwinklig zueinanderstehenden Linien (Achsen) besteht. Die waagerechte Linie wird als X-Achse oder „Abszisse", die senkrechte Linie als Y-Achse oder „Ordinate" bezeichnet. In ihrem Schnittpunkt liegt der Nullpunkt beider Achsen.

Auf der Abszisse sind die verschiedenen Testwerte (in Klassen zusammengefaßt) in gleichen Abständen aufgetragen. Auf der Ordinate können die zugehörigen Häufigkeiten, d. h. die Zahl der Resultate, die in jede Klasse fallen, abgelesen werden. Um verschieden umfangreiche Verteilungen direkt vergleichen zu können, drückt man die Klassenhäufigkeiten gewöhnlich in Prozenten der Gesamtgruppe (N) aus. (Auf den Tabellen Abb. 1 und 2 sind die prozentualen Häufigkeiten unter f % angegeben).

Die graphische Darstellung ist ein Hilfsmittel zur ersten Analyse eines Verfahrens. Sie ermöglicht außerdem den anschaulichen Vergleich mehrerer Tests.

[1]) Über die Einteilung oder Klassen siehe Anhang Seite 57

Man unterscheidet verschiedene Arten graphischer Darstellungen, von denen für die Veranschaulichung von Testverteilungen im wesentlichen folgende drei zur Anwendung kommen: 1. das Stabdiagramm oder Histogramm, 2. das Häufigkeits-Polygon und 3. die Summenprozentkurve oder Ogive.

Die beiden ersten Arten sollen jetzt am Beispiel des Durchstreichtests gezeigt werden, über die Ogive wird später (S. 49) zu sprechen sein.

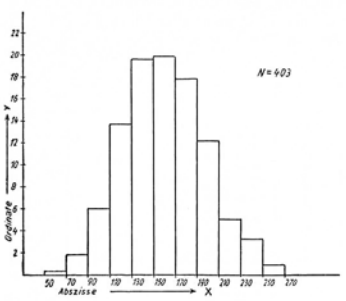

Abb. 3 Stabdiagramm (Histogramm)

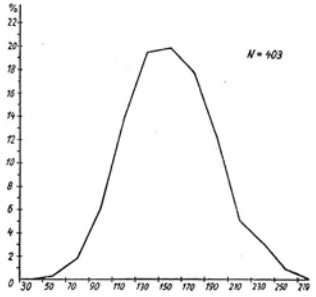

Abb. 4 Häufigkeits-Polygon

Abb. 3 zeigt die Häufigkeitsverteilung des Durchstreichtests als Histogramm. In den Leistungsklassen (X-Achse) sind jeweils 20 verschiedene Werte zusammengefaßt. Darüber sind Säulen errichtet, die anzeigen, wieviel Prozent aller Ergebnisse in die entsprechenden Klassen fallen. (Der Einfachheit halber sind nur die unteren Klassengrenzen angegeben.)

Abb. 4 zeigt die gleiche Verteilung als Häufigkeits-Polygon. Es unterscheidet sich von dem Histogramm dadurch, daß über jeder Klassenmitte nur ein Punkt aufgetragen ist, der die in der Klasse enthaltenen Werte repräsentiert. Die einzelnen Punkte sind durch gerade Linien miteinander verbunden. — An beiden Enden der Leistungsskala (Abszisse) sind Klassen eingetragen, in denen keine Testleistung enthalten ist, die also die Häufigkeit 0 haben. Darin zeigt sich, daß mit dem Test die wirklichen Leistungsextreme erfaßt werden.

Da die zweite Art der Darstellung einfacher und übersichtlicher ist, wird sie im folgenden für die Beispiele verwendet.

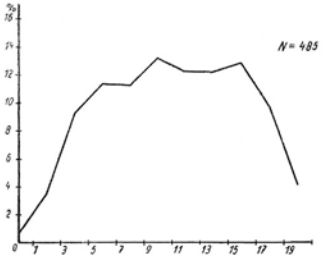

Abb. 5 Analogietest

Abb. 5 veranschaulicht die Häufigkeitsverteilung des *Analogietests* nach den Daten der Tabelle in Abb. 2. Der Vergleich mit der Verteilung des Durchstreichtests (Abb. 4) läßt einige wesentliche Unterschiede erkennen. Trotz gleicher Klassenzahl ist das Polygon hier bedeutend breiter und endet (vor allem bei den hohen Werten) nicht auf der X-Achse. Darin zeigt sich ein Mangel dieses Analogietests: Die extremen Leistungen können nicht erfaßt werden, da unter 0 und über 20 hinaus keine Werte möglich sind. (Grund: zu wenig Einzelaufgaben oder zu lange Bearbeitungszeit.)

Wie die Analyse des Tests ergeben hat, enthielt die hier dargestellte Fassung des Analogietests noch andere „Schönheitsfehler", die vermutlich die Unregelmäßigkeit des Polygons verursachen. (Trotzdem ist diese erste Fassung als Beispiel gewählt worden, da man an ihr die verschiedenen Schwierigkeiten der Testeichung besser als an einer gut ausgefeilten Fassung demonstrieren kann.

An der graphischen Darstellung lassen sich die Eigentümlichkeiten einer Häufigkeitsverteilung ohne umständliche Berechnungen erkennen. Man kann sehen, ob sich die Werte symmetrisch oder unsymmetrisch, gleichmäßig oder ungleichmäßig verteilen, ob sie an einem Ende der Leistungsskala gehäuft auftreten, ob bestimmte Klassen besonders schwach belegt sind usw. Dementsprechend können Schlüsse auf eventuelle Verbesserungsmöglichkeiten des betreffenden Tests gezogen werden.

Zum Vergleich sollen die graphischen Darstellungen einiger anderer Verfahren abgebildet werden, die dem Leser aus dem Text zum Teil bekannt sind, und an denen sich die Mängel in der Konstruktion dieser Verfahren zeigen. Diese mangelhaften Verteilungskurven beziehen sich auf die *Probefassungen* der Tests, die auf Grund der so gewonnenen Aufschlüsse verbessert oder als Gruppentests nicht mehr verwendet werden.

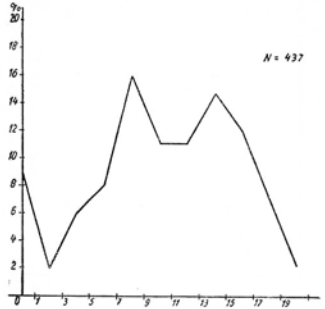

Abb. 6 Telegrammtest

Die Verteilung der Resultate im *Telegrammtest* (s. S. 36) zeigt eine Häufung der Ergebnisse in der untersten Klasse und einen deutlich zweigipfligen Kurvenzug. — Es ist zu vermuten, daß ein Teil der Probanden die gestellte Aufgabe, u. a. die Bedingung, daß die Wortzahl der Lösung begrenzt ist, nicht verstanden hat. (Mangel der mündlichen Instruktion!) — Die Zweigipfligkeit läßt, wie in vielen Fällen, darauf schließen, daß die im Test erfaßte Leistung auf zwei verschiedene, voneinander unabhängige Fähigkeiten zurückgeht, die — würden sie isoliert erfaßt — vermutlich eingipflig streuen würden. Gerade beim Telegrammtest liegt diese Annahme nahe, da zu seiner Lösung neben einer vorwiegend abstrakten Intelligenzleistung (schnelles und sicheres Erfassen eines Sachverhaltes und Herausheben des Wesentlichen) eine erhebliche sprachliche Gewandtheit bzw. Routine im Aufsetzen von Telegrammen erforderlich ist. Es wäre zu spekulativ, weitere Mängel des Tests aus der graphischen Darstellung ablesen zu wollen.

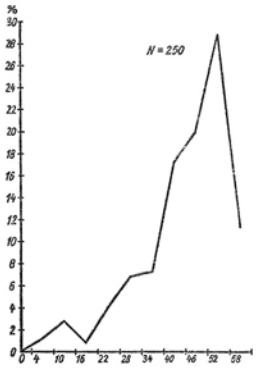

Abb. 7 Lückentest

Bei der in Abbildung 7 dargestellten Häufigkeitsverteilung eines *Lückentests* fällt die starke Rechtsverschiebung des Gipfels auf. Die Werte zeigen eine Häufung bei den guten Leistungen (48 Prozent aller Resultate liegen im obersten Viertel der Leistungsskala), ein Zeichen, daß der Test in dieser Fassung zu leicht ist. (Grund: zu lange Bearbeitungszeit, zu wenig Einzelaufgaben — in diesem Falle Lücken — oder zu geringe Schwierigkeit der Lücken.) — Da die Analyse der Testergebnisse gezeigt hat, daß die guten Leistungen in der Regel mit kürzeren Arbeitszeiten zusammenfallen, kann nur eine Erschwerung der Lücken die Häufigkeitsverteilung verbessern. — Die Unregelmäßigkeit der Kurve läßt noch weitere Konstruktionsmängel des Tests erkennen. Eine quantitative Auswertung dieser Fassung des Lückentests ist also von der Häufigkeitsverteilung her sehr in Frage gestellt.

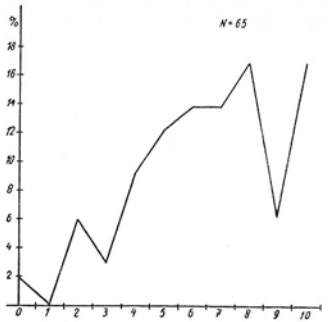

Abb. 8 Sprichwortvergleichtest

In Abb. 8 ist die Häufigkeitsverteilung des *Sprichwortvergleichtests* (s. S. 30) wiedergegeben. Eine Interpretation erübrigt sich nach den vorhergegangenen Beispielen, zumal da dem Leser die Mängel des Tests bereits bekannt sind (vor allem Abhängigkeit der einzelnen Aufgaben voneinander).

Die in Abb. 9 dargestellte Verteilung eines *Buchstabenreihentests* (s. S. 34/II) zeigt das Fehlen des rechten Kurvenastes und eine damit verbundene Rechtslage des Gipfels bei sonst ziemlich ausgeglichenem Kurvenlauf. — Es handelt sich um eine probeweise angewandte Fassung, die nur 15 Aufgaben enthielt. Die Analyse der Ergebnisse ließ erkennen, daß der Test für die gegebene Arbeitszeit zu kurz war. Die Anwendung einer erweiterten Fassung (25 Aufgaben) ergab eine wesentlich günstigere Häufigkeitsverteilung.

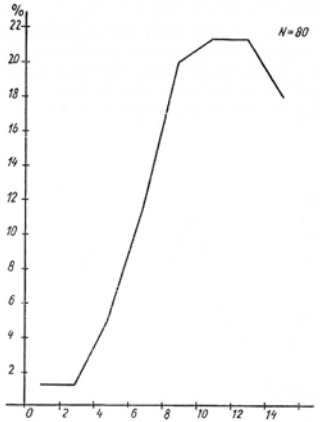

Abb. 9 Buchstabenreihentest

Nicht immer liegen die Mängel, die in einer graphischen Darstellung sichtbar werden, in der Konstruktion des betreffenden Testverfahrens. Auch die Bewertungsmethode kann — wenn sie dem Test nicht angepaßt ist — die Häufigkeitsverteilung negativ beeinflussen. Als Beispiele seien die Darstellungen verschiedener Bewertungen des Analogie- und des Durchstreichtests angeführt.

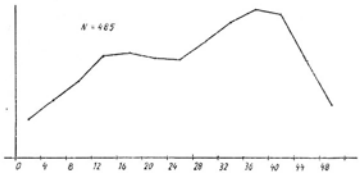

Abb. 10 Analogietest - Leistungspunkte

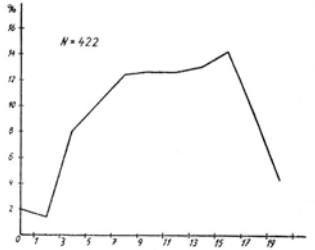

Abb. 11 Analogietest - richtige Lösungen

In Abb. 10 und 11 sind zwei Bewertungsmethoden des Analogietests gegenübergestellt. Der Häufigkeitsverteilung in Abb. 10 liegt ein differenziertes Punktsystem zugrunde, das die einzelnen Aufgaben ihrer Schwierigkeit entsprechend mit verschiedenen Punkten bewertet. Abb. 11 dagegen zeigt die Verteilung der richtigen Lösungen; hierbei wurde nur ausgezählt, wie viele Aufgaben richtig gelöst waren und die Summe davon gebildet. (Bei mehrdeutigen Aufgaben sind halbrichtige Lösungen einbezogen.)

Der Vergleich beider Darstellungen zeigt, daß die einfachere Bewertungsmethode wie in vielen ähnlichen Fällen eine günstigere Verteilung ergibt. Der Einwand, daß durch diese Methode die Beurteilung der Einzelleistungen vergröbert sei, wird durch einen statistischen Vergleich der Resultate beider Methoden widerlegt. Dieser ergibt, daß die relative Stellung der einzelnen Probanden innerhalb der Gesamtgruppe durch die differenzierte Methode nur unwesentliche Veränderungen erfährt.

Bei der Auswertung des Durchstreichtests (s. S. 39) war man vor die Frage gestellt, ob es besser sei, lediglich die Anzahl der richtigen Durchstreichungen zugrunde zu legen oder auch die Zahl der übersehenen Buchstaben, also der Fehler mit einzubeziehen („Fehlerprozent").

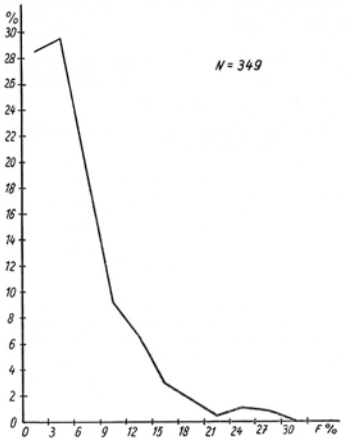

Abb. 12 Durchstreichtest I, Fehlerprozente

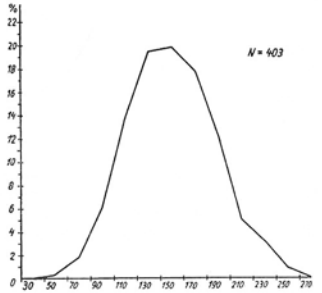

Abb. 13 Durchstreichtest I, richtige Durchstreichungen

In Abb. 12 ist die Häufigkeitsverteilung der Fehlerprozente wiedergegeben. Man kann daran erkennen, daß der größte Teil der Resultate bei den niedrigen Prozentwerten liegt, während nur wenige Probanden in ihren Leistungen mittlere und hohe Fehlerprozente aufweisen (vgl. S. 39). — Abb. 13 veranschaulicht demgegenüber die Verteilung der richtigen Durchstreichungen in der gleichen Teilaufgabe. Es ist deutlich, daß diese Werte für die Begutachtung wesentlich ergiebiger sind.

Die Häufigkeitsverteilung gibt aber nicht nur Hinweise auf die Eigenart oder die praktische Brauchbarkeit eines Verfahrens, sie ist vor allem auch für die Gewinnung der Normen von großer Bedeutung. Wenn sie nämlich bestimmte, jetzt zu besprechende Bedingungen erfüllt, kann man bei der Eichung eines Tests mit statistischen Techniken arbeiten, die die objektive Beurteilung der einzelnen Leistungen sehr erleichtern.

2. *Normalverteilung*

Ein Vergleich der Häufigkeitsverteilungen vieler gut durchgearbeiteter Tests läßt trotz charakteristischer Unterschiede bestimmte gemeinsame Züge erkennen: 1. Die Resultate verteilen sich fortlaufend, ohne deutlich voneinander abgehobene Gruppen über die gesamte Leistungsskala. 2. Die mittleren Klassen sind in der Regel sehr dicht belegt, während nach den Extremen zu die Häufigkeiten stark abfallen. (s. z. B. Abb. 4). Die Form der graphischen Darstellungen erinnert an die aus der Mathematik bekannte Gauß'sche Glockenkurve oder „Normalkurve", die sich ähnlich bei der Messung vieler biologischer, anatomischer und physiologischer Merkmale ergibt (vorausgesetzt, daß ein repräsentativer und genügend großer Querschnitt von Merkmalsträgern zugrunde liegt).

Es hat sich gezeigt, daß sich eine normale und einigermaßen abhebbare psychische Funktion in der Gesamtbevölkerung in der Regel ebenso „normal" verteilt wie viele körperliche Merkmale. Daher ist es gebräuchlich geworden, die Normalkurve als Maßstab für den praktischen Wert eines Tests anzusehen.

An der Glockenkurve sollen jetzt kurz die Voraussetzungen für die bei einer Normalverteilung anwendbaren statistischen Methoden demonstriert werden.

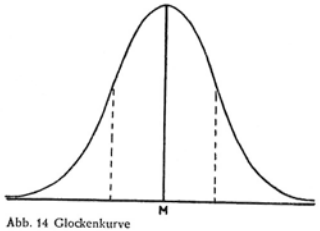

Abb. 14 Glockenkurve

Bei der Betrachtung der nebenstehenden Kurve fallen bestimmte Merkmale ihrer Form auf:

1. Sie ist völlig symmetrisch.

2. Sie hat nur einen Gipfel, der über dem Mittelpunkt der Basislinie liegt.

3. Die Kurvenäste zeigen einen gleichmäßigen, in ihrem Verlauf charakteristischen Abfall nach beiden Seiten vom Gipfelpunkt. Sie gehen nämlich über einem bestimmten Punkt der Basislinie von einem konvexen in einen konkaven Linienzug über.

4. Die Kurve nähert sich der Basislinie immer mehr, ohne sie jedoch zu berühren.

In den äußeren Merkmalen dieser Kurve drücken sich bestimmte mathematische Gesetzmäßigkeiten aus. Sie beziehen sich auf das Verhältnis von Abschnitten der X-Achse zu den darüberliegenden Kurvenflächen. Da die Gesamtfläche einer Kurve den Umfang der in der Verteilung enthaltenen Gruppe repräsentiert, bedeutet die Kenntnis des Zusammenhangs von Basisabschnitten und zugehörigem Flächenanteil, daß man für bestimmte Einheiten der X-Achse mathematisch feststellen kann, wie groß die jeweilige Häufigkeit, d. h. die Zahl der darüber in der Kurvenfläche enthaltenen Fälle ist.

Die Gesetzmäßigkeiten der Normalkurve lassen sich auf jede annähernd normale Häufigkeitsverteilung übertragen. Man bemüht sich darum, einen Test so zu konstruieren, daß seine Resultate eine normale Verteilung aufweisen. Bei einem solchen Test kann man Normen errichten, die die Leistungen der Prüflinge ihrem Stand innerhalb der gesamten Gruppe entsprechend ausdrücken (s. Standardwerte).

Wenn auch psychische Messungen, die an einer begrenzten Zahl von Probanden durchgeführt werden, nie zu idealen Glockenkurven führen, so gibt es doch bestimmte statistisch kontrollierbare Grenzen, innerhalb derer eine Häufigkeitsverteilung als normal behandelt werden kann. Man hat verschiedene Möglichkeiten, diese Grenzen festzustellen:

a) Einen ersten Eindruck liefert bereits der Vergleich eines Häufigkeitspolygons oder Histogramms mit der Normalkurve, durch den man vor allem die Symmetrie und die Regelmäßigkeit der Verteilung feststellen kann. In manchen Fällen reicht dieser Vergleich schon aus, um mangelnde Übereinstimmung erkennen zu lassen. (Vgl. z. B. die Abbildungen 6—9 mit Abb. 4!)

b) Genauer und einfacher ist die Prüfung mit dem „Wahrscheinlichkeitsnetz", eine andere graphische Methode, auf die hier nur hingewiesen werden kann.

c) Hat sich beim graphischen Vergleich eine Verteilung als annähernd normal erwiesen, so kann mathematisch festgestellt werden, wie stark die Verteilung von der Normalverteilung abweicht, indem man den Grad der *Schiefheit* (Abweichung von der Symmetrie) und den Grad der *Flachheit* (Abweichung von der normalen Höhe) berechnet.

d) Eine andere sehr genaue statistische Methode ist das χ^2 *(chi)*-Verfahren (PEARSON), bei dem die einzelnen Klassenhäufigkeiten mit den entsprechenden Häufigkeiten einer Normalverteilung, mit ihren sogenannten „Erwartungswerten" verglichen werden.

Die einzelnen Methoden können hier nicht durchgeführt werden; jedes statistische Lehrbuch gibt darüber Auskunft. — Bei der Prüfung der Häufigkeitsverteilungen des Durchstreichtests zeigte es sich, daß die richtigen Durchstreichungen in allen vier Teilaufgaben ausreichend normal verteilt waren, so daß bei der Eichung dieses Tests die gebräuchlichen statistischen Techniken angewandt werden können. Dagegen hat die hier als Beispiel angeführte Fassung des Analogietests keine normale Verteilung ergeben.

C.

Gewinnung von Normen

Wenn die Häufigkeitsverteilung eines Tests bekannt ist, kann mit der Eichung im engeren Sinne, also mit der Errichtung von Bewertungsnormen begonnen werden.

Es gibt zwei Arten von Normwerten, mit denen man Testleistungen vergleichbar ausdrücken kann: Standardwerte und Prozentränge (s. S. 17). Beide werden aus der Verteilung der Testergebnisse einer großen und charakteristischen Gruppe von Probanden gewonnen. *Standardwerte* entsprechen gleich großen Leistungsunterschieden; sie eignen sich darum vor allem für statistische Vergleiche und Kombinationen. *Prozentränge* entsprechen gleich großen Gruppen von Prüflingen; man kann an ihnen ablesen, wo die Leistung eines Prüflings innerhalb seiner Berufs-, Alters- oder Bildungsgruppe liegt.

1. *Standardwerte*

Ein Standardwert gibt an, wie weit die Leistung eines Menschen über oder unter der normalen Durchschnittsleistung liegt, mit anderen Worten: er zeigt den Abstand einer Testleistung vom Mittelwert der gesamten Verteilung. Auf die Berechnung der Standardwerte kann hier nicht eingegangen werden, da zu ihrem Verständnis die Kenntnis bestimmter statistischer Begriffe erforderlich ist. Im Anhang (S. 61) ist diese Berechnung am Beispiel des Durchstreichtests durchgeführt. Die einzelnen Standardwerte werden zu einer Skala zusammengestellt. An ihr liest der Auswerter mühelos den Wert für jede Testleistung ab.

Da für alle Tests mit normaler Häufigkeitsverteilung solche Standardskalen bestehen, kann der Gutachter die Ergebnisse verschiedener Verfahren direkt miteinander vergleichen und z. B. feststellen, auf welchem Fähigkeitsgebiet ein Prüfling die besten, auf welchem Gebiet er weniger gute Leistungen erzielt hat. Weiter kann untersucht werden, ob von einem Prüfling auf Grund seines Begabungsniveaus einheitliche oder sehr unterschiedliche Leistungen zu erwarten sind.

Ein anderer, sehr wichtiger Vorteil der Standardwerte gegenüber den absoluten Testwerten ist die Möglichkeit, sie mathematisch zusammenzufassen.

Hierfür einige Beispiele:

Um im Durchstreichtest die Gesamtleistung eines Prüflings in einem Wert auszudrücken, kann man die Standardwerte der einzelnen Teilaufgaben addieren und ihre Summe durch

4 teilen, d. h. man kann das arithmetische Mittel der einzelnen Standardwerte berechnen. Diese Berechnung ist in der folgenden Tabelle für einige Probanden (s. S. 40 Urliste) durchgeführt. (Die Werte beziehen sich auf die Zahl der richtigen Durchstreichungen.)

Lfd. Nr.	I		II		III		IV		Mittel der Standardwerte
	R	St	R	St	R	St	R	St	
1	138	45	108	42	38	42	47	50	45
2	137	44	175	61	42	44	53	54	51
3	163	52	149	54	37	42	45	49	49
4	176	55	173	61	62	56	43	48	55
5	203	63	152	54	49	49	31	41	52

usw.

R = richtige Durchstreichung; St = Standardwert; I—IV = Nr. der Teilaufgabe.

Es hat sich gezeigt, daß im Durchstreichtest die Gesamtwerte (s. letzte Spalte der Tabelle) charakteristischer und zuverlässiger sind als die Ergebnisse der einzelnen Teilaufgaben.

Oft ist es zweckmäßig, die Ergebnisse eines Prüflings in den Intelligenztests (evtl. getrennt nach theoretischer und praktischer Intelligenz) den Ergebnissen in den Arbeitstests gegenüberzustellen. Dem Personalreferenten wird damit die Auslese nach Leistungsgesichtspunkten erleichtert, besonders wenn er von einer großen Zahl von Bewerbern nur einen geringen Teil einstellen kann und sehr verschiedene Stellen zu besetzen hat. Mit Hilfe der aus den Standardwerten kombinierten Gesamtwerte kann er eine grobe Vorauslese treffen und anschließend an Hand der differenzierten Eignungsbeurteilungen die endgültige Entscheidung fällen.

Auch die Zusammenfassung *sämtlicher* Testleistungen eines Prüflings in *einer* Zahl ist mit Hilfe der Standardwerte möglich. Wenn diese Zahl für das berufliche Leistungsniveau charakteristisch sein soll, muß mit differenzierten statistischen Methoden die Bedeutung der verwendeten Tests für das betreffende Berufsbild nachgewiesen worden sein; denn die Berufe stellen ja sehr verschiedene Anforderungen an die Fähigkeiten der Bewerber. Bei einer Kombination der Standardwerte muß man den einzelnen Tests dementsprechend verschiedenes Gewicht beilegen.

Sind z. B. Sekretärinnen und Stenotypistinnen mit der gleichen Testserie geprüft worden, wie es bei kleineren Prüfgruppen aus praktischen Gründen oft geschieht, so kann die Bedeutung der einzelnen Tests dabei verschieden sein. Dem Ergebnis eines Intelligenztests wird bei der Begutachtung einer selbständigen Sekretärin mehr Gewicht beizulegen sein als bei der Begutachtung einer Stenotypistin, die vorwiegend mechanische Schreibarbeit leisten soll. Dagegen wird ein Test, der die mechanische Sorgfalt oder die Gleichmäßigkeit und Ausdauer bei der Arbeit erfaßt, im Berufsbild der Stenotypistin mehr Gewicht haben. Solche Abstufungen sind aber nur dann zulässig, wenn die Gültigkeit der einzelnen Tests für die verschiedenen Berufe statistisch exakt bestimmt und nicht mehr oder weniger willkürlich geschätzt worden ist.

Derartige Kombinationen der Standardwerte zu einem Gesamtwert sind jedoch immer nur ein Hilfsmittel, das für eine grobe Vorauslese geeignet ist; eine differenzierte Beurteilung lassen sie nicht zu. Man muß — um beraten und richtig einsetzen zu können — immer auch die Stärken und Schwächen eines Prüflings in ihrem Zueinander berücksichtigen und darum für die einzelnen mit den Tests angegangenen Fähigkeiten getrennte Werte angeben.

2. *Prozentränge*

Ein Prozentrang drückt aus, wieviel Prozent aller *Probanden* einer bestimmten Bevölkerungsgruppe mit ihren Leistungen unter einem Testwert liegen. Wenn also ein Bewerber für den mittleren Verwaltungsdienst in einem Test den Prozentrang 50 (P_{50}) erreicht hat, so

bedeutet es, daß seine Testleistung besser ist als die Leistungen von 50 Prozent der mittleren Beamten, an denen der Test geeicht worden ist. Prozentskalen müssen vor allem für solche Tests angelegt werden, die keine normale Häufigkeitsverteilung aufweisen und für die darum Standardwerte nicht berechnet werden können. Aber auch sonst ist es gebräuchlich, neben den Standardwerten Prozentränge für die einzelnen Testleistungen anzugeben, da sie für den Praktiker leichter zu interpretieren sind.

Die Berechnung der Prozentränge ist im Anhang (S. 62) am Beispiel des Durchstreichtests durchgeführt. Sie geht von einer „summierten Häufigkeitsverteilung" aus, d. h. von einer Verteilung, bei der die einzelnen Klassenhäufigkeiten (f) von der niedrigsten Klasse ansteigend schrittweise addiert sind, so daß die Häufigkeit der obersten Klasse gleich der Gesamtzahl der Fälle ist.

Wenn die einzelnen Häufigkeiten in Prozente der Gesamtgruppe umgerechnet werden, kann eine solche Verteilung als Summenprozentkurve oder Ogive graphisch dargestellt werden. Der praktische Wert einer Ogive besteht darin, daß man an ihr den Prozentrang für jeden beliebigen Wert einer Verteilung direkt ablesen kann und umgekehrt bestimmen kann, unter welchen Wert der Leistungsskala ein bestimmter Prozentsatz aller Resultate fällt. Denn jeder Punkt der Kurve repräsentiert die Zahl der Probanden (in %), deren Resultate unter einem bestimmten Testwert liegen.

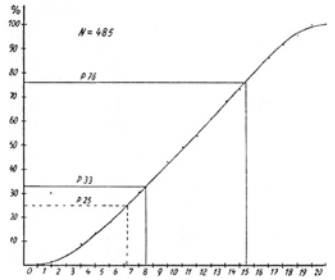

Abb. 15 Ogive für den Analogietest

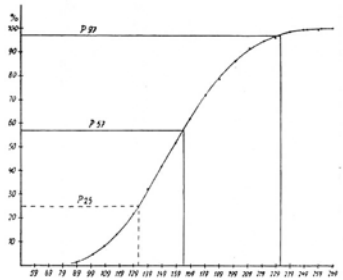

Abb. 16 Ogive für den Durchstreichtest

In Abb. 15 und 16 sind die Summenprozentkurven für den Analogie- und den Durchstreichtest wiedergegeben. Wie beim Histogramm und Polygon sind die Leistungsklassen auf der X-Achse und die Häufigkeiten auf der Y-Achse abgetragen. Da sich die summierten Häufigkeiten auf die oberen Klassengrenzen beziehen, sind sie jeweils über den oberen Klassengrenzen als Punkte eingezeichnet. Die einzelnen Punkte sind durch einen Kurvenzug verbunden. (Dabei sind kleinere Unregelmäßigkeiten ausgeglichen worden.) Die Ogive beginnt also über der unteren Klassengrenze der niedrigsten besetzten Klasse mit 0 Prozent und endet über der oberen Grenze der höchsten besetzten Klasse mit 100 Prozent.

An der Kurve in Abb. 15 kann man z. B. sehen, daß im Analogietest einer Testleistung von 8 richtigen Lösungen der Prozentrang 33 und einer Testleistung von 15 richtigen Lösungen der Prozentrang 76 entspricht. — Im Durchstreichtest (Abb. 16) haben 165 richtige Durchstreichungen den Prozentrang 57, 232 Durchstreichungen den Prozentrang 97.

Will man bei einem Test diejenigen Prüflinge bestimmen, die ihren Leistungen entsprechend zum untersten Viertel der gesamten Berufsgruppe gehören, so kann man an der Kurve feststellen, unter welchem Testwert 25 Prozent aller Resultate liegen. Für den Analogietest ist das der Wert 7, d. h. alle Prüflinge, die 7 und weniger richtige Lösungen haben, gehören in diesem Test zum untersten Viertel ihrer Berufsgruppe. — Im Durchstreichtest ist der entsprechende Wert 133, also alle Prüflinge, die 133 und weniger richtige Buchstaben gestrichen haben, fallen unter die 25-Prozent-Grenze.

Prozentränge können an Hand der Summentafel auch mathematisch bestimmt werden (s. S. 62). Wenn jedoch eine Häufigkeitsverteilung infolge zu kleiner Probandenzahl oder aus anderen Gründen nicht ganz gleichmäßig ist, wird die Ogive zur Aufstellung der Prozentskala herangezogen, da sie weniger als die Summentafel durch die zufällige Zusammensetzung der Gruppe oder durch kleinere Mängel in der Konstruktion oder Durchführung eines Tests beeinflußt wird.

Prozentskalen können nach Berufs- oder Altersgruppen getrennt angelegt werden, dementsprechend läßt sich die Leistung eines Prüflings an den für ihn zutreffenden Normen messen.

Für den Durchstreichtest, bei dessen Auswertung die Altersunterschiede der Prüflinge deutlich in Erscheinung treten, wird im folgenden eine Auswertungstabelle wiedergegeben, an der man für die einzelnen Altersstufen die einem Testergebnis entsprechenden Prozentränge ablesen kann. (Die angegebenen Werte beziehen sich auf die Klassenmitten.)

Testwerte Klassen-Grenzen	Prozentränge						
	20—24	25—29	30—34	35—39	40—44	45—49	50 u. m.
260—269	99						
250—259	98	99					
240—249	96	98	99				
230—239	92	97	98	99			
220—229	88	94	96	97	99	99	
210—219	84	89	93	94	96	98	99
200—209	79	84	88	90	93	95	98
190—199	73	78	82	85	88	91	96
180—189	65	70	75	79	83	86	92
170—179	57	61	66	70	75	79	87
160—169	47	51	56	60	66	70	81
150—159	37	41	45	50	56	59	73
140—149	25	31	34	40	44	47	64
130—139	14	20	24	30	33	35	53
120—129	7	12	15	20	23	25	41
110—119	3	6	8	12	15	17	30
100—109	1	3	4	6	9	11	15
90— 99		1	2	3	4	6	9
80— 89			1	1	1	3	3
70— 79						1	1
60— 69							

Abb. 17 Auswertungstabelle (Altersnormen) Durchstreichtest I

Die in Abb. 17 gezeigte Auswertungstabelle für den Durchstreichtest I läßt z. B. erkennen, daß für ein Resultat von 185 richtigen Durchstreichungen ein 20jähriger Prüfling P_{65}, ein 25jähriger Prüfling P_{70}, ein 30jähriger P_{75} und schließlich ein 50jähriger Prüfling P_{92} erhält. Das gleiche Testergebnis kann also bei einem Prüfling leistungsmäßig eben über dem Durchschnitt liegen (P_{65}), während es bei einem anderen eine ganz hervorragende Leistung darstellt (P_{92}). Diese großen Unterschiede werden auch einen skeptischen Leser davon überzeugen, daß die Leistungen der älteren Prüflinge nicht ungerecht beurteilt werden. (Ja, man darf sogar betonen, daß die Kenntnis der Altersnormen allein eine wirklich gerechte Beurteilung gewährleistet.)

Eine Auswertungstabelle wird für jeden Test angelegt; sie ermöglicht es, ohne umständliche Berechnungen den Prozentrang für jede einzelne Leistung anzugeben. Mit diesen Normen arbeitet der Gutachter bei der Beurteilung.

Prozentränge dürfen zum Unterschied von Standardwerten nicht mathematisch kombiniert werden, da die zwischen ihnen liegenden Leistungsabstände verschieden groß sind. Z. B. besteht bei einer Normalverteilung zwischen den Prozenträngen 50 und 69 der gleiche Leistungsabstand wie zwischen den Prozenträngen 93 und 97. Bei einer Zusammenfassung würden also die extremen Leistungen zu sehr nivelliert.

V.

Korrelationsrechnung und ihre Anwendungsmöglichkeiten bei der Entwicklung von Testverfahren

In den beiden vorhergehenden Abschnitten über die Konstruktion und Eichung von Gruppentests ist die Entwicklung einzelner Verfahren dargestellt worden, bis zu dem Zeitpunkt, an dem die Leistung eines Prüflings nach feststehenden Normen bewertet werden kann. — Um die so entstandenen Tests in eine Eignungsprüfung einschalten zu können, sind jetzt noch folgende Aufgaben zu erfüllen: 1. Die Tests müssen auf ihre Zuverlässigkeit und Gültigkeit geprüft werden, 2. die geeigneten Verfahren müssen zu einer Testserie kombiniert werden, die den Anforderungen des betreffenden Berufsbildes gerecht wird.

Die statistische Methode, welche zur Klärung dieser Fragen beiträgt, ist die „Korrelationsrechnung". Sie ist ein unentbehrliches Hilfsmittel der psychologischen Forschung, da sie nicht nur zur Kontrolle und Kombination von Testverfahren, sondern auch zur Untersuchung vieler praktischer und theoretischer Probleme herangezogen werden kann.

A.

Bedeutung der Korrelationsrechnung

Zunächst ein einfaches Beispiel:

Bei der Eignungsprüfung einer Gruppe von Schreibkräften sind mehrere gut ausgearbeitete Tests durchgeführt worden. Darunter sind zwei, welche die Sorgfalt bei vorwiegend mechanischer Tätigkeit erfassen. — Wenn man die Ergebnisse in diesen beiden Tests vergleicht, so zeigen diejenigen Prüflinge, die in einem Test hohe Werte erreicht haben, auch in dem anderen Test ähnlich hohe Werte. Das gleiche gilt für die Prüflinge mit mittleren und niedrigen Werten. (Träfe das bis zu einem bestimmten Grade nicht zu, so könnten aus den Testresultaten keinerlei Schlüsse auf eine gemeinsame Fähigkeit gezogen werden.)

Vergleicht man nun weiter die Ergebnisse eines dieser Sorgfaltstests mit denen eines Tests zur Prüfung der Sprachgewandtheit, so werden sie kaum Übereinstimmungen zeigen. Hohe Werte im Sorgfaltstest können mit hohen, mittleren oder niedrigen Werten im Sprachtest verbunden sein oder umgekehrt.

Im ersten Fall dieses Beispiels, beim Vergleich der Leistungen zweier Sorgfaltstests, kann man sagen, daß diese beiden Tests einen Zusammenhang, eine Verwandtschaft zeigen, daß sie miteinander „korreliert" sind. Im zweiten Fall, beim Vergleich von Sorgfalts- und Sprachleistung, liegt dagegen keine Korrelation vor.

Es gibt verschiedene Arten der Korrelationsrechnung, von denen bei der Entwicklung und Kontrolle eines Tests vor allem die *Maßkorrelation* zur Anwendung kommt. In einer Maßkorrelation wird der Zusammenhang zwischen zwei Gruppen *gemessener* Werte festgestellt. Sie führt zur Berechnung eines zahlenmäßigen Ausdrucks, des Korrelationskoeffizienten r, der anzeigt, wie eng der Zusammenhang zwischen beiden Wertreihen ist.

Ein Korrelationskoeffizient kann zwischen + 1,00 und — 1,00 liegen. Wenn eine Gruppe von Probanden durch zwei Tests in etwa der gleichen Weise aufgeteilt wird, wenn also die

gleichen Probanden in beiden Tests hohe, mittlere oder niedrige Werte erzielt haben, so spricht man von einer „positiven" Korrelation. Sie ist um so höher, je mehr sich der Koeffizient + 1,00 nähert; bei r = + 1,00 besteht völlige Übereinstimmung in der Reihenfolge und im relativen Abstand der einzelnen Probanden. Schneiden umgekehrt diejenigen Probanden, die in einem Test gute Leistungen gezeigt haben, im anderen Test schlecht ab, so liegt eine „negative" Korrelation vor. Sie ist um so höher, je mehr sich der Koeffizient — 1,00 nähert. — Der Zusammenhang zwischen zwei Wertreihen ist also um so geringer, je näher die Korrelation an 0,00 liegt.

In der Testpraxis sind extrem hohe Korrelationen sehr selten, die Koeffizienten zwischen zwei Verfahren liegen in der Regel zwischen 0,20 und 0,70, die Zuverlässigkeitskoeffizienten zwischen 0,70 und 0,90.

Bei r = 0,00 bis 0,20 besteht kein oder nur ein unwesentlicher Zusammenhang; r = 0,20 bis 0,40 zeigt einen Zusammenhang, der jedoch gering ist; r = 0,40 bis 0,70 läßt einen wesentlichen Zusammenhang erkennen, und r = 0,70 bis 1,00 zeigt eine hohe bis sehr hohe Korrelation.

Für die Interpretation eines Korrelationskoeffizienten muß man wissen, daß er keinen Prozentsatz von irgend etwas ausdrückt. Also r = 0,80 bedeutet nicht, daß bei 80% der Probanden die beiden Werte ihrer Stellung nach übereinstimmen oder etwa, daß die Güte der Korrelation 80% beträgt. Man kann Korrelationskoeffizienten auch nicht in dem Sinne vergleichen, daß z. B. 0,80 einen viermal so engen Zusammenhang darstellte wie 0,20; denn die Beziehung zwischen den beiden Wertreihen steigert sich erheblich mehr als die zahlenmäßige Größe des Korrelationskoeffizienten.

Die Berechnung von r ist im Anhang (S. 64) durchgeführt, ihr methodischer Ansatz soll im folgenden kurz dargestellt werden, damit auch der Leser, den die technischen Einzelheiten weniger interessieren, sich ein Bild von dieser so wichtigten statistischen Methode machen kann.

Zunächst wird eine *Korrelationstabelle* angelegt, die zeigt, wie oft jeder Wert der einen Messung mit jedem Wert der anderen Messung zusammentrifft. (Wie in den Häufigkeitstabellen werden die Werte einer umfangreichen Reihe dabei zu Klassen zusammengefaßt.)

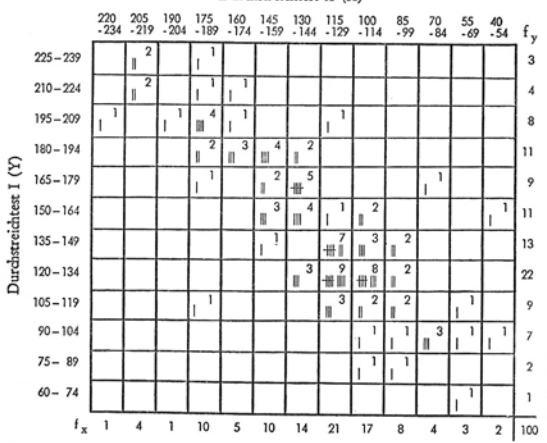

Abb. 18
Korrelationstabelle (Strichliste) für Durchstreichtest I und II

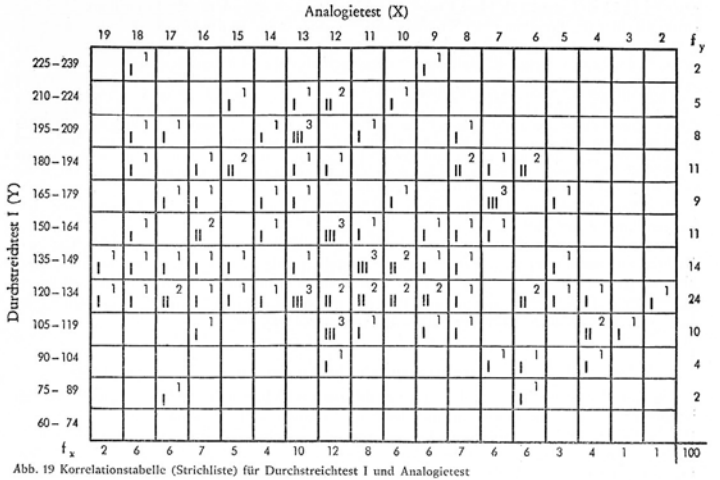

Abb. 19 Korrelationstabelle (Strichliste) für Durchstreichtest I und Analogietest

Die Wertskala des einen Tests (X) wird dabei auf der waagerechten Achse, die des anderen Tests (Y) auf der senkrechten Achse abgetragen. Für jeden Probanden wird ein Strich in dem Feld der Korrelationstabelle gemacht, in welchem sich die beiden seinen Testresultaten entsprechenden Reihen schneiden.

Wenn z. B. ein Proband in der ersten Teilaufgabe des Durchstreichtests 235 Buchstaben und in der zweiten Teilaufgabe 211 Buchstaben richtig gestrichen hat, so werden diese Leistungen durch einen Strich im zweiten Feld der obersten Reihe symbolisiert (s. Abb. 18).

Die Summen der Häufigkeitswerte jeder waagerechten und senkrechten Reihe sind in den Spalten f_x und f_y eingetragen; sie ergeben die Häufigkeitsverteilungen beider Tests.

Man kann schon aus der Verteilung der Striche in einer Korrelationstabelle einen Eindruck davon gewinnen, ob zwischen den beiden Wertreihen ein Zusammenhang besteht und wie eng er etwa ist. Je eindeutiger diagonal eine Verteilung ist, desto enger ist der Zusammenhang.

Im ersten Beispiel (Abb. 18) läßt die Anordnung der Striche auf eine relativ hohe Korrelation schließen, während die sehr verstreute Verteilung im zweiten Beispiel (Abb. 19) darauf hinweist, daß zwischen den Leistungen beider Tests kein — oder nur ein unwesentlicher — Zusammenhang besteht. (Da der Leser weiß, um welche Verfahren es sich in beiden Tabellen handelt, wird er über diesen Unterschied nicht erstaunt sein.) Der Korrelationskoeffizient für das erste Beispiel beträgt 0,80, für das zweite Beispiel 0,21.

An der hohen Korrelation zwischen den beiden Teilaufgaben des Durchstreichtests ist zu erkennen, daß sie weitgehend die gleiche Funktion erfassen. Es macht also im allgemeinen für die geprüfte Fähigkeit keinen Unterschied aus, ob in dem Test zwei oder drei Buchstaben gestrichen worden sind. — Die sehr geringe Korrelation zwischen Durchstreich- und Analogietest dagegen zeigt, daß die in beiden Verfahren erfaßten Fähigkeiten zum großen Teil voneinander unabhängig sind, denn Korrelationen um 0,20 und niedriger können nur bei sehr großer Probandenzahl als statistisch gesichert angesehen werden.

In bestimmten Fällen kann die Maßkorrelation nicht angewandt werden, so z. B. wenn für den Vergleich keine exakt gemessenen Werte vorliegen oder wenn die Zahl der Probanden zu

gering ist (30 und weniger). Hier kann unter Umständen eine andere sehr brauchbare Methode Auskunft über das Bestehen eines Zusammenhanges geben, nämlich die *Rangkorrelation*. Beispiel:

Die Gültigkeit eines neu entwickelten Intelligenztests soll geprüft werden. Man hat diesen Test mit den Teilnehmern mehrerer Verwaltungskurse durchgeführt. Um die Testergebnisse mit den im Unterricht gezeigten intellektuellen Leistungen vergleichen zu können, ließ man die Teilnehmer durch die Lehrkräfte auf die Höhe ihrer Intelligenz hin beurteilen. Da es sich um getrennte Kurse mit verschiedenen Lehrkräften handelte, war ein Vergleich sämtlicher Probanden nicht ohne weiteres möglich. In solchem Fall hilft man sich mit der Aufstellung einer „Rangreihe" für jeden Lehrgang; die Teilnehmer werden in eine Reihe gebracht und ihren Rangplätzen entsprechend numeriert. (Bei einem Kursus von 20 Teilnehmern erhält also der intelligenteste die Zahl 1, der am wenigsten intelligente die Zahl 20). Diese Rangreihen werden dann mit den Testleistungen korreliert. Im Anhang S. 66 ist ein Beispiel für eine solche Korrelation gegeben.

Eine Rangkorrelation führt zu einem Korrelationskoeffizienten (ρ), dessen Bedeutung derjenigen von r entspricht.

Mit Hilfe der Korrelationsrechnung kann der Grad des Zusammenhanges zwischen zwei Reihen gemessener oder geschätzter Werte festgestellt werden. Der Korrelationskoeffizient sagt aber nichts über die *Ursachen* eines bestehenden Zusammenhanges. Diese Ursachen zu untersuchen, ist Aufgabe der psychologischen Forschung, die sich seit langem mit der Feststellung der in den Tests wirksamen Faktoren befaßt (Faktorenanalyse).

B.
Anwendung der Korrelationsmethode bei der Entwicklung von Tests

Die Korrelationsmethode wird in den verschiedensten Gebieten der angewandten Psychologie herangezogen. Im Zusammenhang mit der Entwicklung von Tests braucht man sie vor allem für die Kontrolle der Gültigkeit und der Zuverlässigkeit eines Verfahrens, für die Zusammenstellung von Testserien und für die Bestimmung des relativen Gewichts der einzelnen Tests innerhalb der gesamten Serie.

1. Kontrolle der Gültigkeit

Der Prüfer muß sicher sein, daß sich die von ihm entwickelten Tests für den bestimmten Zweck — also etwa für die Prüfung von Bewerbern des mittleren Verwaltungsdienstes — eignen. Darum muß die Gültigkeit dieser Verfahren festgestellt werden. Hierfür gibt es zwei Möglichkeiten:

a) Man korreliert die Testergebnisse mit denen eines bereits geeichten Tests, der die gleiche Fähigkeit prüft.

b) Man korreliert die Testergebnisse mit bestimmten fachlichen Leistungen (Examensnoten, meßbare Berufserfolge, Beurteilungen durch Vorgesetzte, Lehrer usw.).

Daß man auch die Gültigkeit der Einzelaufgaben auf diese Weise prüfen kann, wurde bereits erwähnt (S. 22).

Von den erprobten Tests werden diejenigen in die Testserie aufgenommen, welche die höchste Gültigkeit gezeigt haben.

2. Kontrolle der Zuverlässigkeit

Der Prüfer muß sich auf die Ergebnisse seiner Tests verlassen können. Er muß also wissen, in welchem Ausmaß sie durch zufällige äußere Faktoren beeinflußbar sind. Zur Feststellung der Zuverlässigkeit gibt es drei Wege:

a) Die Testwiederholung, bei der ein Test zweimal mit den gleichen Probanden durchgeführt wird — die Ergebnisse beider Messungen werden korreliert;

b) der statistische Vergleich der Testwerte mit denen eines Paralleltests;

c) die Testhalbierung.

Bei dieser Methode, auf deren technische Einzelheiten nicht näher eingegangen werden kann, zeigt der Korrelationskoeffizient nur die Zuverlässigkeit des halben Tests. Es läßt sich jedoch mathematisch bestimmen, wie hoch er bei Verlängerung der Tests sein würde (Brown-Spearmansche Formel).

3. Zusammenstellung von Testserien

Bei einer Eignungsprüfung können nicht beliebige Tests angewandt werden, sondern man muß sich bei der Auswahl von Testverfahren nach den Anforderungen der zu prüfenden Berufe richten. Dazu muß die Bedeutung jedes einzelnen Verfahrens für das betreffende Berufsbild korrelationsstatistisch nachgewiesen werden. Es ist wichtig, in einer Testserie solche Verfahren zusammenzustellen, die untereinander niedrig, mit den fachlichen Leistungen jedoch möglichst hoch korreliert sind. Denn einerseits setzen die beruflichen Anforderungen sehr verschiedene Fähigkeiten voraus, die auch mit verschiedenen Tests erfaßt werden müssen, andererseits muß aber eine Testprüfung aus praktischen Gründen möglichst wenig Zeit in Anspruch nehmen.

Mit Hilfe einer gut zusammengestellten Testserie lassen sich mit einer bestimmten Wahrscheinlichkeit Voraussagen über die berufliche Tüchtigkeit eines Bewerbers machen, die um so sicherer sind, je enger der Zusammenhang zwischen Testserie und Berufsanforderungen ist.

4. Bestimmung des relativen Gewichts für die einzelnen Tests

Um die Standardwerte der verschiedenen Tests zu einem Gesamtwert kombinieren zu können, müssen die Tests — entsprechend ihrer Bedeutung innerhalb der Testserie — mit verschiedenem Gewicht belegt werden. Die Bedeutung eines Tests hängt davon ab, wie weit er den Berufsanforderungen entgegenkommt (Gültigkeit) und wie weit er sich von den übrigen Tests der Serie abhebt. Dementsprechend werden die Standardwerte jedes Tests vor der Kombination mit einer konstanten Zahl multipliziert, die dem relativen Gewicht des Tests entspricht.

Es hat sich gezeigt, daß eine statistisch genau ausgewogene Testserie mit der beruflichen Leistung höher korreliert ist als die in ihr enthaltenen Einzeltests, daß also die Gültigkeit der gesamten Serie größer ist als die jedes einzelnen Tests für sich.

VI.
Zusammenfassung und Schluß

Die Psychologie hat eine Reihe von Methoden entwickelt und wendet eine Vielfalt von Verfahren an, um die ihr gestellten praktischen Aufgaben erfüllen zu können. Tests sind zwar nicht die einzige Form, in der psychologische Untersuchungen durchgeführt werden, aber es dürfte nur selten eine solche Untersuchung geben, bei der sie gänzlich fehlen. Insbesondere ist keine Eignungsprüfung denkbar, die nicht in ausgedehntem Maße von der Testmethode Gebrauch macht.

Überall dort aber, wo Tests und Testserien eingeschaltet sind oder gar die Hauptsubstanz einer Prüfung ausmachen, ist die genaue Kenntnis ihrer Wirkungsweise unerläßlich.

Einige Autoren haben zum Ausdruck gebracht, daß jeder gut entwickelte Test gleichsam ein kleines „Kunstwerk" darstellt: in der Leichtigkeit und Klarheit seiner Form und Anwendung läßt er nicht mehr erkennen, welche langwierigen Vorarbeiten zu seiner Entwicklung nötig sind. Das verführt den Laien sehr leicht dazu, den Test bisweilen für eine müßige „Spielerei" zu halten. Nichts wäre verkehrter als das!

Es war die Aufgabe der vorliegenden Arbeit, in großen Zügen aufzuzeigen, welche wissenschaftlichen und technischen Voraussetzungen der Testmethode zugrunde liegen. Es wird jedem Leser klar geworden sein, daß eine am Schreibtisch beliebig ersonnene Aufgabe alles andere ist als ein Test. Nur dort, wo solche Aufgaben gewissenhaft an der Realität überprüft und in ihrer Anwendung und Deutung aufs sorgfältigste auf die wirklichen Verhältnisse abgestimmt sind, geht es an, das Verfahren als Test zu bezeichnen.

Bevor wir einen Test in der Praxis verwenden, müssen wir wissen: *was* prüft der Test? Prüft er jene Eigenschaften, die wir zu bestimmen beabsichtigen? Wie weit können wir sicher sein, daß wir den gewünschten Funktionsbereich mit ihm ansprechen? Allein die Bestimmung dieser seiner Gültigkeit erfordert die Anwendung hochentwickelter statistischer Methoden.

Bevor wir einen Test in der Praxis verwenden, müssen wir ferner ermittelt haben: wie zuverlässig ist er im Gebrauch? Sind die Ergebnisse, die wir mit ihm von einem bestimmten Menschen erhalten, heute die gleichen wie morgen? Gibt der Test uns also Resultate, die nicht mehr oder weniger zufällig, sondern wirklich für diesen Menschen charakteristisch sind? Wenn die Ergebnisse eines Tests schwanken und unzuverlässig sind, können wir überzeugt davon sein, daß ein solcher Test *gar nichts* von Bedeutung ermittelt. Die Zuverlässigkeit der vorhandenen Tests ist recht verschieden. Sie muß durch entsprechende Methoden festgestellt und dem Prüfer bekannt sein.

Ebenfalls *vor* Anwendung eines Tests müssen seine Normen durch sorgfältige, ausgedehnte und repräsentative Vorversuche geklärt und statistisch verarbeitet sein. Ein Test sagt nur dann etwas aus, wenn die Vergleichsbasis der Häufigkeitsverteilung in der allgemeinen Bevölkerung vorliegt. Nur dann wissen wir, ob die Leistung eines Menschen gut oder schlecht ist, wenn uns bekannt ist, wieviel Prozent seiner Alters- oder Berufsgenossen eine schlechtere bzw. eine bessere Leistung zeigen als er. Der einzelnen vorliegenden Leistung selbst könnte niemand mit einiger Sicherheit ansehen, welche Güte sie besitzt.

Es war das Ziel dieser Arbeit, deutlich zu machen, wie ausgedehnt und langwierig die Vorarbeiten sind, die notwendig sind, um ein Prüfverfahren zu einem Testverfahren zu entwickeln, und wie ernsthaft jeder Test der Bewährungsprobe und Kontrolle unterzogen wird, bevor er zur Anwendung gelangt. Wenn sich der Leser die Mühe gemacht hat, die überwiegend statistisch-technischen Schritte bei der Entwicklung eines Testverfahrens auf ihren Sinn hin zu überprüfen, dann wird er auch die Erkenntnis gewonnen haben, daß eine laienhafte Verwendung des Tests alle vorausgegangene Mühe wieder hinfällig machen kann. Nur dem psychologisch geübten Prüfer, der die Wirkungsweise und Grundlagen der von ihm verwandten Testverfahren genau kennt, ist es möglich, aus den Testergebnissen die richtigen Schlüsse zu ziehen!

Testverfahren sind *objektive* Verfahren. Es ist geschildert worden, mit welchen häufig recht zeitraubenden Voruntersuchungen und mit welchen keineswegs einfachen Mitteln die Objektivität der Bewertung gesichert werden muß. Damit soll und kann in entscheidender Weise die subjektive Willkür der einzelnen Beurteiler ausgeschaltet werden. Der Proband braucht nicht mehr in Sorge zu sein, daß er von persönlichen Einstellungen des Prüfers abhängig ist und daß seine Leistung von zufälligen menschlichen Beziehungen her beurteilt wird. Er bewährt sich lediglich der Aufgabe — nicht dem Prüfer gegenüber. Es kann kein Zweifel sein, daß die höchstmögliche Objektivität das Ziel einer jeden Untersuchung sein muß; hier spricht die Verantwortung dem Prüfling gegenüber das entscheidende Wort.

Überall in der Welt haben sich Tests und Testserien — auf dem Gebiet der Intelligenz, der beruflichen Eignung, der Schulleistung, der Persönlichkeit usw. — bewährt. Unzählige Arbeiten, vor allem des Auslandes, zeigen die praktische Gültigkeit der erzielten Resultate. Diese Erfolge sind nicht zufällig und nicht leicht gewonnen. Sie haben eine mühsame und sorgfältige Aufbauarbeit der einzelnen Verfahren zur Voraussetzung: alles das, was eine Aufgabe zu dem macht, was wir *Test* nennen dürfen.

Als Ergänzung zum Hauptteil dieser Schrift sind im folgenden für diejenigen Leser, die sich näher mit den Fragen der Testeichung befassen wollen, einige technische Methoden zusammengestellt. Diese Zusammenstellung erhebt keinen Anspruch auf Vollständigkeit; viele wichtige technische Techniken können nicht berücksichtigt werden, da zu ihrem Verständnis mehr als Schulkenntnisse der Mathematik erforderlich sind.

Auch die hier angeführten Methoden werden nicht im einzelnen beschrieben oder mathematisch abgeleitet; darüber ist in statistischen Lehrbüchern von berufenen Autoren ausführlicher geschrieben worden, als es im Rahmen dieser einführenden Arbeit möglich wäre.

Statistischer Anhang

1. Einteilung von Klassen (S. 41, 42, 52)

Wenn man mit einer umfangreichen Reihe von Werten statistisch arbeiten will, faßt man immer mehrere aufeinanderfolgende Werte zu *Klassen* zusammen. Die Klassen werden bezeichnet durch Angabe der Klassengrenzen oder der Klassenmitten. Für die *Klassengrenzen* nimmt man am einfachsten ganze Zahlen, im Durchstreichtest z. B. 60—69, 70—79, 80—89. Häufig findet man auch die Angaben 59,5—69,5, 69,5—79,5, 79,5—89,5 usw., die vielleicht mathematisch exakter, aber weniger übersichtlich sind. (Es ist üblich, Testergebnisse wie *kontinuierliche* Merkmalsreihen zu behandeln, bei denen jeder Wert als Abstand innerhalb einer stetigen Zahlenfolge aufgefaßt wird. Der Testwert 7 z. B. umfaßt den Bereich 6,5—7,5, der Testwert 60 den Bereich 59,5—60,5.) Die *Klassenmitte* ist der Mittelwert (M) beider Klassengrenzen, für die Klasse 60—69 z. B. (wenn wir sie so definieren, daß sie von 59,5 bis 69,5 reicht) $\frac{59{,}5 + 69{,}5}{2} = 64{,}5$. Statistischen Berechnungen werden grundsätzlich die Klassenmitten zugrunde gelegt, sie repräsentieren sämtliche in einer Klasse enthaltenen Resultate.

Der von einer Klasse umfaßte Abschnitt der Wertskala wird als *Klassenbreite* oder *Intervall* (i) bezeichnet. Die Größe des Intervalls hängt ab von der „Variationsbreite" des gemessenen Merkmals, d. h. von der Differenz zwischen dem höchsten und niedrigsten auftretenden Wert, und von der jeweils zu wählenden Klassenanzahl. Sie errechnet sich nach der einfachen Beziehung

$$i = \frac{\text{Variationsbreite}}{\text{Klassenanzahl}},$$

wobei man die nächstliegende abgerundete Zahl der Reihe, 2, 3, 5, 10, 15 oder 20 wählt.

Die Anzahl der Klassen soll nicht unter 10 und nicht über 20 betragen.

2. Berechnung des arithmetischen Mittels (M)

Die einzelnen Testwerte einer Gruppe von Probanden können zu einem *Mittelwert* zusammengefaßt werden. Dieser Mittelwert repräsentiert in gewisser Hinsicht die gesamte Gruppe. Es gibt verschiedene Arten von Mittelwerten, unter denen der bei der Eichung von Tests gebräuchlichste das arithmetische Mittel (M) ist — in der Umgangssprache als „Durchschnitt" bezeichnet. M ist die Zahl, die man erhält, wenn man die Summe der zu mittelnden Werte (X) durch ihre Anzahl (N) teilt, also

$$M = \frac{\Sigma X}{N}$$

Man kann M auch mit der Häufigkeitstabelle berechnen, was vor allem bei großen Gruppen üblich ist. In diesem Fall multipliziert man die verschiedenen Werte bzw. Klassenmitten (X) mit den dazugehörigen Häufigkeiten (f), bildet die Summe dieser Produkte und dividiert durch die Summe der Häufigkeitszahlen (N).

1	2	3	4
Klassen-grenzen	X	f	fX
230—239	234,5	1	234,5
220—229	224,5	2	449,0
210—219	214,5	3	643,5
200—209	204,5	8	1636
190—199	194,5	8	1556
180—189	184,5	5	922,5
170—179	174,5	7	1221,5
160—169	164,5	6	987
150—159	154,5	7	1081,5
140—149	144,5	6	867
130—139	134,5	15	2017,5
120—129	124,5	9	1120,5
110—119	114,5	6	667
100—109	104,5	5	522,5
90— 99	94,5	1	94,5
80— 89	84,5	2	169
70— 79	74,5	0	0
60— 69	64,5	1	64,5
		N = 92	14274,0

Beispiel 1: Berechnung von M für den Durchstreichtest (Teilgruppe „gehobener Dienst") an Hand der Häufigkeitstabelle.

Die einzelnen Schritte:

1. Jede Klassenmitte (X) wird mit der zugehörigen Häufigkeit (f) multipliziert. Die Ergebnisse werden unter fX in die Häufigkeitstabelle eingetragen (Spalte 4).
2. Man bildet die Summe dieser Produkte ($\Sigma fX = 14274{,}0$).
3. Diese Summe wird durch die Gesamtzahl der Werte (N = 92) dividiert, denn

$$M = \frac{\Sigma fX}{N}$$

Man erhält $\frac{14274{,}0}{92} = 155{,}15$ richtige Durchstreichungen als Durchschnittsleistung der Bewerber für den gehobenen Verwaltungsdienst im Durchstreichtest I.

Diese Methode ist wegen der erheblichen Rechenarbeit, besonders wenn sie ohne Rechenmaschine durchgeführt werden muß, sehr umständlich. Es wird deshalb zur Berechnung von M oft mit einer wesentlich vereinfachten Methode, der sog. *Kurzmethode* gearbeitet. Bei der Kurzmethode geht man von einem „angenommenen" Mittelwert (A) aus; zweckmäßig nimmt man dafür eine Klassenmitte, die etwa im Zentrum der Gesamtverteilung liegt. Dann berechnet man die Korrektur, um die der angenommene Mittelwert verändert werden muß, wenn man den wirklichen Mittelwert erhalten will. Auf Grund mathematischer Ableitungen gilt für die Korrektur die Formel:

$$c = \frac{\Sigma fx'}{N}$$

Darin bedeuten:

f die verschiedenen Klassenhäufigkeiten,

x' die „Abweichungen" der Klassenmitten vom angenommenen Mittelwert (A) in Einheiten der Klassenbreite. Die Klasse, in der A liegt, hat die Abweichung 0, die nächsthöhere Klasse + 1, die darauffolgende + 2 usw. (positive Abweichungen); die unter A liegenden Klassen haben die Abweichungen —1, —2 usw. (negative Abweichungen). Die Klassen werden also zunächst behandelt, als hätten sie die Größe 1.

N die Summe aller Häufigkeiten.

Um die Korrektur für A zu erhalten, muß der Ausdruck c anschließend mit der Klassenbreite (i) multipliziert werden. Für den wirklichen Mittelwert gilt dann die Beziehung

$$M = A + i\,\frac{\Sigma fx'}{N} \quad \text{oder} \quad M = A + i\ c$$

Beispiel 2: *Berechnung von M für den Durchstreichtest (Teilgruppe „gehobener Dienst") mit der Kurzmethode.*

	1	2	3	4
	X	f	x'	fx'
	234,5	1	+ 8	+ 8
	224,5	2	+ 7	+ 14
	214,5	3	+ 6	+ 18
	204,5	8	+ 5	+ 40
	194,5	8	+ 4	+ 32
	184,5	5	+ 3	+ 15
	174,5	7	+ 2	+ 14
	164,5	6	+ 1	+ 6
A =	154,5	7	0	0
	144,5	6	— 1	— 6
	134,5	15	— 2	— 30
	124,4	9	— 3	— 27
	114,5	6	— 4	— 24
	104,5	5	— 5	— 25
	94,5	1	— 6	— 6
	84,5	2	— 7	— 14
	74,5	0	— 8	0
	64,5	1	— 9	— 9
		N=92		$\Sigma fx' = +6$

Die einzelnen Schritte:

1. An der Häufigkeitstabelle wird ein Mittelwert geschätzt (A = 154,5).
2. Die Abweichungen der Klassenmitten (X) vom angenommenen Mittel (A) werden in die Tabelle eingetragen (Spalte 3).
3. Jede Abweichung (x') wird mit der zugehörigen Häufigkeit (f) multipliziert, dabei müssen die Vorzeichen beachtet werden (Spalte 4).
4. Die Summe dieser Produkte wird gebildet ($\Sigma fx' = 6$.)
5. Diese Summe wird durch die Summe der Häufigkeiten (N = 92) dividiert und mit der Klassenbreite (i = 10) multipliziert. Das Ergebnis ist die Korrektur für A (im Beispiel
$$10 \cdot \frac{6}{92} = 0{,}65).$$
6. Die Korrektur wird zu A addiert.

Der wirkliche Mittelwert ist also

$$M = 154{,}5 + 0{,}65 = 155{,}15.$$

Der Vergleich mit der ersten Methode (Beispiel 1) zeigt, daß das gleiche Ergebnis hier mit wesentlich einfacheren Berechnungen gewonnen worden ist.

Da ein Mittelwert die Leistungen einer Gruppe als Ganzes charakterisiert, kann er zum Vergleich verschiedener Gruppen herangezogen werden. In bestimmten Fällen ist das arithmetische Mittel hierfür weniger geeignet, so z. B. wenn es durch wenige ausgefallene Werte zu stark in einer Richtung beeinflußt würde oder wenn die Häufigkeitsverteilung an ihren Enden sog. offene Klassen enthält, z. B. "120 und mehr", "59 und weniger". Hier wird ein anderer Mittelwert, der sog. Zentralwert (Z) benutzt. Über die Berechnung von Z wird unter 5) zu sprechen sein.

3. Berechnung der Standardabweichung (σ)

Der Mittelwert allein genügt jedoch nicht zur Kennzeichnung einer Gruppe von Werten, sondern es müssen daneben auch noch Angaben über die „Streuung" gemacht werden.

Ein Streuungsmaß zeigt an, wie stark die einzelnen Werte von ihrem Mittelwert abweichen, wie groß also die individuellen Unterschiede innerhalb einer Gruppe sind. Bei gleichem Mittelwert sind die Leistungen zweier Gruppen sehr verschieden zu beurteilen, wenn die Streuungsmaße stark differieren. — Das gebräuchlichste und zuverlässigste Streuungsmaß ist die *Standardabweichung* (σ); sie faßt die Abweichung aller Werte von arithmetischen Mittel in einem Wert zusammen. Die Standardabweichung kann nur bei ausreichend normaler Häufigkeitsverteilung berechnet werden. Die einfachste Formel, mit der bei kleinen Gruppen gearbeitet wird, lautet:

$$\sigma = \pm \sqrt{\frac{\Sigma\ x^2}{N - 1}}$$

Dabei ist x die Abweichung jedes einzelnen Wertes vom Mittelwert, also x = X — M. Die beiden Vorzeichen drücken aus, daß die Standardabweichung nach beiden Seiten vom Mittelwert aus gewertet werden muß.

Bei großen Gruppen wird σ an Hand der Häufigkeitsverteilung berechnet und zwar nach der Formel:

$$\sigma = \pm \sqrt{\frac{\Sigma\ f\ x^2}{N - 1}}$$

Hier ist x die Abweichung jeder einzelnen Klassenmitte vom arithmetischen Mittel.

Beispiel 3: Berechnung von σ für den Durchstreichtest (Gesamtgruppe „mittlerer Verwaltungsdienst") an der Häufigkeitstabelle.

1	2	3	4	5
X	f	x	fx	fx^2
264,5	1	+ 104,96	+ 104,96	11 016,60
254,5	2	+ 94,96	+ 189,92	18 034,80
244,5	4	+ 84,96	+ 339,84	28 872,81
234,5	9	+ 74,96	+ 674,64	50 571,01
224,5	7	+ 64,96	+ 454,72	29 538,61
214,5	13	+ 54,96	+ 714,48	39 267,82
204,5	21	+ 44,96	+ 944,16	42 449,43
194,5	28	+ 34,96	+ 978,88	34 221,64
184,5	31	+ 24,96	+ 773,76	19 313,05
174,5	41	+ 14,96	+ 613,36	9 175,87
164,5	39	+ 4,96	+ 203,36	1 008,67
154,5	41	— 5,04	— 206,64	1 041,47
144,5	39	— 15,04	— 586,56	8 821,86
134,5	40	— 25,04	— 1001,60	25 080,06
124,5	35	— 35,04	— 1226,40	42 973,06
114,5	20	— 45,04	— 900,80	40 572,03
104,5	14	— 55,04	— 770,56	42 411,62
94,5	10	— 65,04	— 650,40	42 302,02
84,5	5	— 75,04	— 375,20	28 155,01
74,5	2	— 85,04	— 170,08	14 463,60
64,5	1	— 95,04	— 95,04	9 032,60
	403			538 323,64

Bei den Daten der nebenstehenden Häufigkeitstabelle (Spalte 1 und 2) handelt es sich um die gleiche Verteilung wie in der Strichliste Abb. 1 (S. 41) des Hauptteils.

Die einzelnen Schritte:

1. Die Abweichungen der Klassenmitten (X) vom arithmetischen Mittel (M = 159,54) werden berechnet (Spalte 3).
2. Die einzelnen Abweichungen (x) werden mit den zugehörigen Häufigkeiten (f) multipliziert (Spalte 4).
3. Diese Produkte (fx) werden mit den zugehörigen Abweichungen (x) multipliziert, man erhält die Abweichungsquadrate (Spalte 5).
4. Die Summe der Abweichungsquadrate wird gebildet ($\Sigma fx^2 = 538\,323{,}64$) und in die Formel für σ eingesetzt.

Also: $\sigma = \pm \sqrt{\frac{538323{,}64}{402}} = \pm\ 36{,}59$

Auch die Berechnung von σ kann durch die Kurzmethode wesentlich vereinfacht werden. Bei der Eichung von Tests wird mit dieser Methode gearbeitet, da mit ihr M und σ gleichzeitig gewonnen werden können. Die Formel für σ nach der Kurzmethode lautet:

$$\sigma = \pm\ i \sqrt{\frac{\Sigma\ fx'^2}{N - 1} - \left(\frac{\Sigma\ fx'}{N}\right)^2} \quad \text{oder} \quad \sigma = \pm\ i \sqrt{\frac{\Sigma\ fx'^2}{N - 1} - c^2}$$

Beispiel 4: *Berechnung von M und σ für den Durchstreichtest (Gesamtgruppe „mittlerer Verwaltungsdienst") mit der Kurzmethode*

1	2	3	4	5
x	f	x'	fx'	fx'²
264,5	1	+10	+ 10	100
254,5	2	+ 9	+ 18	162
244,5	4	+ 8	+ 32	256
234,5	9	+ 7	+ 63	441
224,5	7	+ 6	+ 42	252
214,5	13	+ 5	+ 65	325
204,5	21	+ 4	+ 84	336
194,5	28	+ 3	+ 84	252
184,5	31	+ 2	+ 62	124
174,5	41	+ 1	+ 41	41
A=164,5	39	+ 0	0	0
154,5	41	— 1	— 41	41
144,5	39	— 2	— 78	156
134,5	40	— 3	—120	360
124,5	35	— 4	—140	560
114,5	20	— 5	—100	500
104,5	14	— 6	— 84	504
94,5	10	— 7	— 70	490
84,5	5	— 8	— 40	320
74,5	2	— 9	— 18	162
64,5	1	—10	— 10	100
	N=403		—200	5482

Die einzelnen Schritte:

1. Ein angenommenes Mittel wird festgesetzt (A = 164,5).
2. Die Abweichungen der einzelnen Klassen von A werden in die Tabelle eingetragen (Spalte 3).
3. Jede Abweichung (x') wird mit der zugehörigen Häufigkeit (f) multipliziert (Spalte 4).
4. Die Produkte (fx') werden mit x' multipliziert — man erhält so die Abweichungsquadrate (Spalte 5).
5. Die Summe aller Abweichungen ($\sum fx' = -200$) wird berechnet (Spalte 4 unten).
6. Die Summe der Abweichungsquadrate ($\sum f x'^2 = 5492$) wird berechnet (Spalte 5 unten).
7. Um c zu erhalten, wird die Summe der Abweichungen ($\sum fx'$) durch die Gesamtzahl der Häufigkeiten (N) dividiert, also
$$c = \frac{-200}{403} = -0{,}496$$
8. Die errechneten Werte werden in die Formel für σ eingesetzt und ergeben im vorliegenden Fall:
$$\sigma = \pm 10 \sqrt{\frac{5482}{402} - (-0{,}496)^2} = \pm 36{,}59$$

Gleichzeitig erhält man den Wert für M nach der obenstehenden Formel (s. S. 58)

$$M = 164{,}5 + 10 \cdot -0{,}496 = 159{,}54$$

Um die repräsentative Bedeutung der Standardabweichung zu verstehen, muß man sich die Gesetzmäßigkeiten der Normalkurve (s. S. 46) vor Augen zu führen. Bei einer Normalkurve bestehen konstante Beziehungen zwischen bestimmten Basisabschnitten und den darüberliegenden Flächenanteilen. Für die Einheiten von σ, die vom Mittelpunkt der Basislinie aus nach beiden Seiten gemessen werden, sind die zugehörigen Anteile der Kurvenfläche (in Prozenten der Gesamtfläche ausgedrückt) tabellarisch festgehalten. Daran kann man z. B. ablesen, daß dem Abschnitt 1,00 σ 34,13% der Gesamtfläche entsprechen, dem Abschnitt ± 1,00 σ also 68,26% der Gesamtfläche.

Auf das vorliegende Beispiel mit M = 159,54 und σ = ± 36,59 angewandt, bedeutet das: im Durchstreichtest liegen — Normalverteilung vorausgesetzt — gut ⅔ aller Leistungen im Wertbereich von 159,54 ± 36,59, d. h. zwischen 123 und 196 richtigen Durchstreichungen.

In gleicher Weise läßt sich für jede andere Einheit der Standardabweichung, etwa für 1,50 σ, 2,00 σ 3,00 σ usw. der zugehörige Flächenanteil ablesen. Man kann an der Häufigkeitstabelle kontrollieren, ob die empirische Verteilung diesen „Erwartungswerten" entspricht und hat damit einen Maßstab dafür, ob es sich um eine normale Häufigkeitsverteilung handelt.

Im Durchstreichtest (Beispiel 4) liegen die mittleren ⅔ aller Werte zwischen 123 und 197 richtigen Durchstreichungen, ein Zeichen dafür, daß der Mittelwert (M = 159,54) und das Streuungsmaß (σ = 36,59) die Gesamtgruppe „mittlerer Verwaltungsdienst" repräsentieren und zum Vergleich mit anderen Gruppen herangezogen werden können.

Außer ihrer Bedeutung für die Beurteilung einer Gruppe von Werten, z. B. der Testleistungen einer bestimmten Berufsgruppe, bildet die Standardabweichung die Grundlage für die Gewinnung der Standardwerte, von der im nächsten Abschnitt zu sprechen sein wird.

4. Berechnung der Standardwerte (S. 11, 55—57)

Ein Standardwert (St) ist ein allgemein vergleichbares Maß für den Abstand eines Testresultates (X) vom Mittelwert (M) der gesamten Verteilung. Um den Standardwert für eine Testleistung zu bestimmen, muß zunächst die Abweichung des Testwertes vom arithmetischen Mittel (X — M = x) durch die Standardabweichung (σ) dividiert werden. Die Größe $\frac{x}{\sigma}$ drückt aus, um wieviel Einheiten der Standardabweichung ein Testwert über oder unter dem Durchschnitt liegt (positives oder negatives Vorzeichen). Da bei statistischen Berechnungen die Dezimalstellen und die negativen Vorzeichen hinderlich sind, ist es üblich, die $\frac{x}{\sigma}$ — Werte mit 10 zu multiplizieren und das Ergebnis zu 50 zu addieren. Dementsprechend berechnet man Standardwerte nach der Formel

$$St = 10 \cdot \frac{x}{\sigma} + 50$$

Eine Testleistung, die im Mittelwert der gesamten Verteilung liegt (x=0), hat also den Standardwert 50, eine Testleistung, die um 1 σ über dem Mittelwert liegt, hat den Standardwert 60, eine Testleistung, die um 1 σ unter dem Mittelwert liegt, den Standardwert 40 usw.

Beispiel 5: *Berechnung der Standardwerte für einige Testleistungen im Durchstreichtest I (Gesamtgruppe „mittlerer Verwaltungsdienst")*

Im Durchstreichtest I ist M = 159,54 und σ = ± 36,59 (s. Beispiel 4, S. 60)

Testwert	$\frac{x}{\sigma}$	Standardwert	
199	$\frac{199 - 159{,}54}{36{,}59} = +1{,}08$	$10 \cdot +1{,}08 + 50 = 60{,}8$	abger. 61
136	$\frac{136 - 159{,}54}{36{,}59} = -0{,}64$	$10 \cdot -0{,}64 + 50 = 43{,}6$	abger. 44
78	$\frac{78 - 159{,}54}{36{,}59} = -2{,}23$	$10 \cdot -2{,}23 + 50 = 27{,}7$	abger. 28

Die auf diese Weise berechneten Standardwerte verschiedener Tests können unmittelbar verglichen und statistisch kombiniert werden. An bestimmten Tabellen kann für jede beliebige Einheit der Standardabweichung der dazugehörige prozentuale Anteil der Normalkurve abgelesen werden. Man kann also feststellen, wieviel Prozent aller Fälle bei einer Normalverteilung unter oder über einem bestimmten Testwert liegen.

Für die Werte im Beispiel 5:

199 Durchstreichungen liegen 1,08σ über dem Mittelwert; 1,08σ entsprechen 35,99% der Verteilung, M + 1,08 σ also 85,99%, d. h. die Leistung ist besser als rund 86% aller Leistungen.
136 Durchstreichungen liegen 0,64 σ unter dem Mittelwert; 0,64 σ entsprechen 23,89% der Verteilung, M — 0,64 σ also 26,11%, die Leistung ist also besser als rund 26% aller Leistungen.
78 Durchstreichungen liegen nach den gleichen Berechnungen über 1,29% der Leistungen, also fast an der untersten Leistungsgrenze.

Die hier beschriebenen Standardwerte sind nicht mit den sog. T-Werten zu verwechseln. T-Werte drücken die Testleistungen einer „normalisierten" Häufigkeitsverteilung aus, d. h. einer ursprünglich nicht ausreichend normalen Verteilung, die mit bestimmten statistischen Techniken in eine Normalverteilung umgewandelt wurde.

5. Berechnung der Prozentränge (S. 10, 57—59)

Ein Prozentrang (P) bezieht sich auf die Stellung eines Prüflings innerhalb der Verteilung, er drückt aus, wieviel Prozent aller Testleistungen unter einem bestimmten Wert liegen. Zu seiner Berechnung geht man von der *summierten* Häufigkeitsverteilung (Kumulativverteilung) aus.

Im Beispiel 6 ist die summierte Häufigkeitsverteilung für den Durchstreichtest (Gesamtgruppe „mittlerer Verwaltungsdienst") wiedergegeben. Die Spalten 1 und 2 enthalten die gleiche Häufigkeitsverteilung wie die Strichliste Abb. 1 (S. 41) des Hauptteils. Die einzelnen Klassenhäufigkeiten (f) sind — von der niedrigsten Klasse aufsteigend — schrittweise addiert (Spalte 3). Die so gebildeten summierten Häufigkeiten (F) sind in Prozente der Gesamtsumme umgerechnet (Spalte 4). Für die Klasse 220—229 z. B. ist die summierte Häufigkeit 387; das sind $\frac{387 \cdot 100}{403} = 96\%$ aller in der Verteilung enthaltenen Werte.

Die Prozentwerte der Spalte 4 bilden die Grundlage für die Errichtung der Summenprozentkurve oder Ogive, über deren Bedeutung bereits gesprochen wurde (s. S. 49).

Man kann den Prozentrang für ein Testresultat an der Ogive ablesen oder aus der summierten Häufigkeitsverteilung durch Interpolation berechnen.

Beispiel 6: Berechnung des Prozentranges für eine Leistung von 195 richtigen Durchstreichungen im Durchstreichtest 1.

1	2	3	4
Klassengrenzen	f	F	F %
260—269	1	403	100
250—259	2	402	99,75
240—249	4	400	99,25
230—239	9	396	98,25
220—229	7	387	96
210—219	13	380	94,25
200—209	21	367	91
190—199	28	346	85,8
180—189	31	318	78,9
170—179	41	287	71,2
160—169	39	246	61,1
150—159	41	207	51,4
140—149	39	166	41,3
130—139	40	127	31,6
120—129	35	87	21,7
110—119	20	52	13
100—109	14	32	8
90— 99	10	18	4,5
80— 89	5	8	2
70— 79	2	3	0,75
60— 69	1	1	0,25

Die einzelnen Schritte:

1. Eine summierte Häufigkeitsverteilung wird angelegt (s. nebenstehende Tabelle).
2. Man teilt die Häufigkeit (f) der Klasse, in der der betreffende Wert liegt (190—199), durch die Klassenbreite (i = 10) und erhält die Zahl der Werte, die auf jede Einheit (Variante) innerhalb der Klasse kommen, also $^{28}/_{10} = 2{,}8$ Werte.
3. Der Testwert 195 liegt von der unteren Klassengrenze (189,5) um 5,5 Einheiten entfernt. Mithin hat er 5,5 · 2,8 = 15,4 Häufigkeiten mehr als die untere Klassengrenze.
4. Diese Häufigkeit muß zur summierten Häufigkeit der unteren Klassengrenze addiert werden. Die Summe aller Häufigkeiten bis zum Wert 195 beträgt dann 318 + 15,4 = 333,4.
5. Die Zahl wird in Prozente der Gesamtgruppe (N = 403) umgerechnet, also
$$\frac{333{,}4 \cdot 100}{403} = 82{,}73\,\%$$
P_{83} ist also der Prozentrang für den Testwert 195, d. h. 83 % der Gruppe „mittlerer Verwaltungsdienst" liegen mit ihren Leistungen unter 195 richtigen Durchstreichungen.

An der Summentafel kann umgekehrt berechnet werden, unter welchem Wert der Leistungsskala ein bestimmter Prozentsatz aller Testresultate liegt. Man fragt also etwa: Wie sieht die Leistung aus, die gerade von 50 % der Probanden geschafft wird? Die Berechnung dieser Werte, die auch als „Perzentilen" bezeichnet werden, erfolgt nach der Formel

$$P_p = u + \left(\frac{\frac{pN}{100} - F_u}{f_p}\right) i$$

Darin ist:

p der Prozentsatz der Gesamtgruppe, unter dem der gesuchte Weg liegt,
u die untere Grenze der Klasse, in der der gesuchte Wert liegt,
N die Gesamtzahl der Fälle,
F_u die summierte Häufigkeit der Klasse unter u,
f_p die Häufigkeit der Klasse, in der der gesuchte Wert liegt und
i die Klassenbreite.

Beispiel 7: Berechnung von Perzentilen im Durchstreichtest 1.

Um zu bestimmen, welcher Testwert von 25 % aller Prüflinge erreicht worden ist, welches Perzentil also dem Prozentrang 25 entspricht, setzt man die erforderlichen Werte an Hand der summierten Häufigkeitsverteilung (s. Tabelle Beispiel 6) in die obenstehende Formel ein. Hier ist

$$\frac{pN}{100} = \frac{25 \cdot 403}{100} = 100{,}75 \qquad P_{25} = 129{,}5 + \left(\frac{100{,}75 - 87}{40}\right) \cdot 10 = 132{,}94$$

Für P_{75} ist nach der gleichen Berechnung

$$\frac{pN}{100} = \frac{75 \cdot 403}{100} = 302{,}25 \qquad P_{75} = 179{,}5 + \left(\frac{302{,}25 - 287}{31}\right) \cdot 10 = 184{,}42$$

25 % aller Prüflinge erreichen also 133 richtige Durchstreichungen, 75 % erreichen 184 richtige Durchstreichungen in der gegebenen Zeit. (Vergleiche die entsprechenden Werte der Ogive Abb. 16, S. 49.)

6. Berechnung des Zentralwertes (Z) und der Quartil-Abweichung (Q)

Der Zentralwert ist derjenige Punkt in einer Wertskala, der die Eigenschaft hat, daß die Hälfte aller Fälle über und die andere Hälfte unter ihm liegt. Wenn man also bei einem Test sämtliche Resultate ihrer Größe nach ordnet, ist Z der mittelste Wert dieser Reihe. (Ist N eine gerade Zahl, so liegt Z zwischen den beiden mittleren Werten.) Z entspricht dem Prozentrang 50 (P_{50}), da unter ihm 50 % aller Leistungen liegen. Er wird in der eben dargestellten Weise berechnet.

Beispiel 8: Berechnung des Zentralwertes für den Durchstreichtest 1 (Gesamtgruppe „mittlerer Verwaltungsdienst").

$$\frac{pN}{100} = \frac{50 \cdot 403}{100} = 201{,}5 \qquad Z\,(P_{50}) = 149{,}5 + \left(\frac{201{,}5 - 166}{41}\right) \cdot 10 = 158{,}2$$

Der Zentralwert wird gelegentlich zur Kennzeichnung kleinerer Gruppen verwendet; statistischen Berechnungen wird immer das arithmetische Mittel zugrunde gelegt.

Wenn man eine Häufigkeitsverteilung in vier gleich große Gruppen unterteilt, erhält man drei Quartile (Viertelwert-Abstände). Q_1 ist derjenige Punkt der Leistungsskala, unter dem das unterste Viertel aller Werte liegt, also $Q_1 = P_{25}$. Q_2 ist identisch mit dem Zentralwert oder P_{50}. Q_3 bezeichnet den Punkt der Leistungsskala, unter dem drei Viertel aller Werte liegen, also $Q_3 = P_{75}$. Die *Quartil-Abweichung*, die bisweilen als Streuungsmaß berechnet wird, ist die Hälfte des Abstandes zwischen den mittleren 50 Prozent einer Häufigkeitsverteilung, also zwischen den Werten, die P_{25} und P_{75} entsprechen. Für die Quartil-Abweichung (Q im engeren Sinn) gilt die Beziehung

$$Q = \frac{Q_3 - Q_1}{2}$$

Beispiel 9: Berechnung der Quartil-Abweichung für den Durchstreichtest 1.

$Q_1\,(P_{25}) = 132{,}94$ $\quad Q_3\,(P_{75}) = 184{,}42$ (s. Beispiel 7) also

$$Q = \frac{184{,}42 - 132{,}94}{2} = 25{,}74$$

Die Quartil-Abweichung wird gelegentlich zum Vergleich kleinerer Teilgruppen, die sehr extreme Werte enthalten, herangezogen. Sie ist immer kleiner als die Standardabweichung der gleichen Verteilung, darf also nicht direkt mit σ verglichen werden.

7. Berechnung der Maßkorrelation (S. 51—53)

Bei einer Maßkorrelation wird der Zusammenhang zwischen zwei Variablen innerhalb des gleichen Kollektivs, z. B. der Leistungen einer Gruppe von Probanden in zwei verschiedenen Tests, untersucht. Ein Maß für den bestehenden Zusammenhang ist der Korrelationskoeffizient r, für den die Beziehung gilt.

$$r = \frac{\sum xy}{(N-1)\,\sigma_x\,\sigma_y} \qquad \text{(BRAVAIS — PEARSON)}$$

Darin sind

x die Abweichungen der Werte der ersten Variablen von ihrem Mittelwert also $x = X - M_x$

y die Abweichungen der Werte der zweiten Variablen von ihrem Mittelwert also $y = Y - M_y$

σ_x die Standardabweichung der ersten Variablen, also $\sigma_x = \sqrt{\frac{\sum x^2}{N-1}}$

σ_y die Standardabweichung der zweiten Variablen, also $\sigma_y = \sqrt{\frac{\sum y^2}{N-1}}$

Man benutzt sie gelentlich bei kleineren Gruppen (N = 25 und weniger), für die keine Häufigkeitstabelle angelegt wird. Bei umfangreichen Reihen bestimmt man den Korrelationskoeffizienten an Hand einer Korrelationstabelle. Man arbeitet dabei — entsprechend der Kurzmethode zur Berechnung von

M und σ — nicht mit den absoluten Werten, sondern mit ihren Abweichungen von angenommenen Mittelwerten A_x und A_y). Der dadurch entstehende Fehler wird später mathematisch korrigiert. Hier lautet die Formel:

$$r = \frac{\frac{\sum x' y'}{N} - c_x c_y}{\sqrt{\frac{\sum f x'^2}{N} - c_x^2} \sqrt{\frac{\sum f y'^2}{N} - c_y^2}}$$

Darin ist c_x die Korrektur für A_x, also $c_x = \frac{\sum f x'}{N}$ und c_y die Korrektur A_y, also $c_y = \frac{\sum f y'}{N}$

Beispiel 11: *Berechnung der Maßkorrelation zwischen Durchstreichtest I und II an der Korrelationstabelle.*

Die folgende Korrelationstabelle enthält die Leistungen von 100 Probanden (N = 100) im Durchstreichtest I und II. Es handelt sich um die gleichen Daten wie in der Strichliste Abb. 1 (S. 52) des Hauptteils. (Die Striche sind bei der Übertragung weggelassen.)

Durchstreichtest II (X)

Durchstreichtest I (Y)			237	213	197	182	167	152	A_x 121	122	107	92	77	62	47	(1)	(2)	(3)	(4)	(5)	(6)
			+6	+5	+4	+3	+2	+1	0	−1	−2	−3	−4	−5	−6	f_y	y'	fy'	fy'^2	$\Sigma x'$	$\Sigma x'y'$
233	+6			2		1										3	+6	+18	108	+13	79
217	+5			2		1	1									4	+5	+20	100	+15	75
202	+4		1		1	4	1			1						8	+4	+32	128	+23	92
187	+3					2	3	4	2							11	+3	+33	99	+16	48
172	+2					1		2	5				1			9	+2	+18	36	+ 1	2
157	+1							3	4	1	2				1	11	+1	+11	11	− 8	−8
A_y 142	0							1		7	3	2				13	0	0	0	−18	0
127	−1								3	9	8	2				22	−1	−22	22	−31	31
112	−2					1				3	2	2		1		9	−2	−18	36	−15	30
97	−3									1	1	3	1	1		7	−3	−21	63	−28	84
82	−4									1	1					2	−4	− 8	32	− 5	20
67	−5												1			1	−5	− 5	25	− 5	25
(1a)		f_x	1	4	1	10	5	10	14	21	17	8	4	3	2	100		58	660	−42	477
(2a)		x'	+6	+5	+4	+3	+2	+1	0	−1	−2	−3	−4	−5	−6						
(3a)		fx'	6	20	4	30	10	10	0	−21	−34	−24	−16	−15	−12	−42					
(4a)		fx'^2	36	100	16	90	20	10	0	21	68	72	64	75	78	644					
(5a)		$\Sigma y'$	4	22	4	33	18	19	17	−10	−17	−13	−7	−10	−2	58					
(6a)		$\Sigma y'x'$	24	110	16	99	36	19	0	10	34	39	28	50	12	477					

$$\frac{\sum x' y'}{N} = \frac{477}{100} = 4{,}77$$

$$c_x = \frac{-42}{100} = -0{,}42$$

$$c_x^2 = 0{,}18$$

$$c_y = \frac{58}{100} = 0{,}58$$

$$c_y^2 = 0{,}34$$

$$\frac{\sum f x'^2}{N} = \frac{644}{100} = 6{,}44$$

$$\frac{\sum f y'^2}{N} = \frac{660}{100} = 6{,}60$$

Die einzelnen Schritte:

1. Eine Strichliste wird für die beiden zu korrelierenden Tests angelegt (s. S. 52). Die darin enthaltenen Häufigkeiten werden in eine Korrelationstabelle übertragen. (Bei dieser Tabelle wird zweckmäßig am oberen und am linken Rand eine Reihe freigelassen, in die später die Abweichungen x' und y' eingetragen werden können.)
2. Für jede Variable wird an Hand der Verteilungen in den Summenreihen f_x und f_y ein Mittelwert geschätzt. Im Beispiel: $A_x = 137$, $A_y = 142$.
   Die beiden Reihen, in denen A_x und A_y liegen, werden in der Tabelle durch Doppellinien hervorgehoben.
3. Die Abweichungen der einzelnen Klassen beider Variabler (X und Y) von ihrem angenommenen Mittel (A_x bzw. A_y) werden in die Tabelle eingetragen (Spalte 2 und 2a).
4. Jede Abweichung (x' und y') wird mit der zugehörigen Häufigkeit (f_x bzw. f_y) multipliziert (Spalte 3 und 3a).

5. Die Produkte (f_x' und f_y') werden mit den entsprechenden Abweichungen (x' bzw. y') multipliziert (Spalte 4 und 4a).
6. Die Summe der Abweichungen in jeder waagerechten und in jeder senkrechten Reihe ($\Sigma x'$ bzw. $\Sigma y'$) wird berechnet. Hierfür trägt man der besseren Übersicht halber am oberen und am linken Rande der Korrelationstabelle nochmals die in den Spalten 2 und 2a enthaltenen Abweichungen (x' und y') ein, multipliziert die Häufigkeit jedes Feldes einer Reihe mit der zugehörigen Abweichung und berechnet die Summe dieser Produkte für jede waagerechte bzw. senkrechte Reihe. Um z. B. die Summe der Abweichungen für die oberste waagerechte Reihe festzustellen, geht man folgendermaßen vor:

   Man multipliziert die Häufigkeit im ersten belegten Feld (2) mit der zugehörigen Abweichung ($x' = +5$). Zu dem Produkt (10) addiert man das Produkt aus Häufigkeit und Abweichung des nächsten belegten Feldes ($1 \cdot +3 = 3$). Die Summe ($10 + 3 = 13$) wird unter $\Sigma x'$ (Spalte 5) in die Tabelle eingetragen.

   Für die dritte waagerechte Reihe: $(1 \cdot +6) + (1 \cdot +4) + (4 \cdot +3) + (1 \cdot +2) + (1 \cdot -1) = 23$. Die einzelnen Häufigkeiten in den senkrechten Reihen werden dementsprechend mit den Abweichungen y' am linken Rand der Tabelle multipliziert und die Summen dieser Produkte unter $\Sigma y'$ (Spalte 5a) eingetragen. Bei der Summierung ist es wichtig, die negativen Vorzeichen nicht zu übersehen!
7. Die Abweichungssummen $\Sigma x'$ werden mit den zugehörigen y' multipliziert (Spalte 6); die Abweichungssummen $\Sigma y'$ werden mit den zugehörigen x' multipliziert (Spalte 6a).
8. Die einzelnen Werte der Spalten 3 bis 6 und 3a bis 6a werden addiert. Man hat folgende Kontrollmöglichkeiten für die Richtigkeit der Ergebnisse (s. Korrelationstabelle): Die Spalten $\Sigma x'y'$ und $\Sigma y'x'$ (Spalte 6 und 6a) müssen die gleiche Summe ergeben (hier $\Sigma x'y' = 477$), ebenso die Spalten fy' und $\Sigma y'$ (Spalte 3 und 5a, hier $\Sigma fy' = 58$) und die Spalten $\Sigma x'$ und fx' (Spalte 5 und 3a, hier $\Sigma fx' = -42$).

   Die einzelnen Werte und die Gesamtsummen der Spalten fy'^2 und fx'^2 (Spalte 4 und 4a) müssen nachgerechnet werden, da für sie in der Tabelle keine Kontrollen enthalten sind.
9. Aus den erhaltenen Werten wird der Korrelationskoeffizient nach der obenstehenden Formel berechnet. Für das vorliegende Beispiel ergibt sich gemäß den neben der Tabelle stehenden Werten:

$$r = \frac{4{,}77 - (-0{,}42 \cdot 0{,}58)}{\sqrt{6{,}44 - 0{,}18}\ \sqrt{6{,}60 - 0{,}34}} = 0{,}800$$

   Der Korrelationskoeffizient r = 0,80 zeigt also den Grad des Zusammenhanges zwischen der ersten und zweiten Teilaufgabe des Durchstreichtests. (Über die Bedeutung eines Korrelationskoeffizienten s. Hauptteil S. 52.)

Die hier dargestellte Korrelationsmethode wird bei der Untersuchung des Zusammenhangs zwischen zwei Tests in der Regel angewandt. Sie führt aber nur unter bestimmten, vorher zu prüfenden Bedingungen zu zuverlässigen Ergebnissen: 1. Die Zahl der Probanden muß möglichst groß sein (mindestens N = 30). 2. Die Korrelation muß eine binormale Häufigkeitsverteilung aufweisen. Voraussetzung dafür ist die einwandfreie Konstruktion beider Tests und die Güte der in der Korrelation enthaltenen Stichprobe (s. S. 22/23). 3. Zwischen beiden Variablen muß ein „linearer" Zusammenhang bestehen, d. h. die beiden Linien, welche durch die Mittelwerte (M) der waagerechten und senkrechten Reihen einer Korrelationstabelle gezogen werden können (Regressionslinien), müssen annähernd gradlinig verlaufen. (Bei deutlich diagonaler Verteilung erübrigt es sich, die Linearität zu prüfen.)

Sind diese Bedingungen nicht erfüllt, müssen andere Korrelationsmethoden angewandt werden.

8. Berechnung der Rangkorrelation (S. 62)

Die Methode der Rangkorrelation prüft den Zusammenhang zwischen zwei *Rangreihen*, der im Koeffizienten ρ zahlenmäßig ausgedrückt wird. Eine Rangkorrelation wird vor allem dann berechnet, wenn man mit kleinen Gruppen arbeitet oder wenn es sich bei einem bzw. beiden zu vergleichenden Merkmalen um nicht meßbare psychische Qualitäten oder um geschätzte Werte handelt.

Um eine Rangkorrelation durchzuführen, muß man zunächst die Probanden ihren Leistungen entsprechend zu Rangreihen ordnen. Für jeden Probanden wird dann die Differenz (d) seiner beiden Rangplätze festgestellt. Aus der Summe der Quadrate dieser Differenzen (Σd^2) und aus der Gesamtzahl der Fälle (N) wird der Korrelationskoeffizient berechnet nach der Formel

$$\rho = 1 - \frac{6 \sum d^2}{N (N^2 - 1)} \qquad \text{(SPEARMAN)}$$

Beispiel 12: *Berechnung einer Rangkorrelation zwischen zwei Tests.*
Die Leistungen einer Gruppe von Probanden in zwei neuentwickelten Tests sollen auf einen etwa bestehenden Zusammenhang hin untersucht werden. Die Gruppe ist für eine Maßkorrelation zu klein (N = 12). Um einen ersten Überblick zu erhalten, ob die beiden Tests miteinander korreliert sind, wird eine Rangkorrelation durchgeführt.

1	2	3	4	5
Pb.	Test 1	Test 2	d	d^2
A	2,5	4	1,5	2,25
B	8	7	1	1
C	5	6	1	1
D	11	10	1	1
E	10	12	2	4
F	2,5	3	0,5	0,25
G	9	10	1	1
H	12	10	2	4
I	7	5	2	4
J	1	1	0	0
K	4	2	2	4
L	6	8	2	4
N =12				26,50

Die einzelnen Schritte:

1. Die Resultate der Probanden werden für beide Tests der Größe nach geordnet und entsprechend numeriert. Die beste Leistung erhält dabei den Rangplatz 1, die nächstfolgende den Rangplatz 2 usw. Gleichen Resultaten gibt man den Mittelwert der betreffenden Rangplätze. (In Test I z. B. waren die Leistungen der Probanden A und F gleich, sie folgten auf die Leistungen von J. Statt des zweiten und dritten Rangplatzes bekommen hier beide den Platz 2,5.)
2. Name und Rangplätze jedes Probanden (A—L) werden in eine Tabelle eingetragen (Spalte 1—3).
3. Die Differenz (d) beider Rangplätze wird berechnet (Spalte 4).
4. Die Quadrate der Differenzen (d^2) werden gebildet (Spalte 5).
5. Die Summe dieser Quadrate ($\Sigma d^2 = 26{,}50$) wird zu sammen mit N (12) in die Formel eingesetzt.

$$\rho = 1 - \frac{6 \cdot 26{,}5}{12 \cdot 143} = +\,0{,}91$$

Dieses Ergebnis zeigt einen deutlichen Zusammenhang zwischen beiden Tests — wenn es auch wegen der sehr geringen Probandenzahl nur mit Vorbehalt verwendet werden kann.

Die Rangkorrelation zwischen zwei Variablen ist weniger genau als die Maßkorrelation, da sie nicht die absoluten Werte, sondern nur die Reihenfolge der einzelnen Fälle innerhalb der Gruppe berücksichtigt. Sie ist jedoch ein gutes Hilfsmittel, um einen vorläufigen Eindruck vom Bestehen eines Zusammenhanges zu gewinnen, wenn eine Maßkorrelation nicht angewandt werden kann.

Auf einen sehr wichtigen Punkt, der bei jeder statistischen Untersuchung beachtet werden muß, kann hier nur hingewiesen werden, nämlich auf die Prüfung der *Zuverlässigkeit statistischer Ergebnisse.* Sie ist deshalb erforderlich, weil man in der Regel nur mit Ausschnitten *(Stichproben)* aus der Bevölkerungsschicht arbeitet, über die bestimmte Angaben gemacht werden sollen. Die errechneten Werte sind infolgedessen immer nur Näherungen derjenigen Werte, die man bei Zugrundelegen des gesamten Kollektivs erhalten würde.

Bevor man auf Grund eines an einer Stichprobe gewonnenen Ergebnisses gültige Aussagen machen kann, muß man mit bestimmten mathematischen Methoden seinen Genauigkeitsgrad feststellen. Hierzu berechnet man den *Standardfehler* des betreffenden Wertes, mit dessen Hilfe man die Grenzen festlegt, innerhalb derer ein Ergebnis, z. B. ein Mittelwert, ein Streuungsmaß, ein Korrelationskoeffizient oder eine Differenz schwanken kann. Nur wenn dieser Schwankungsbereich eine bestimmte Höchstgrenze nicht übersteigt, kann ein Ergebnis als statistisch gesichert angesehen werden.

Eine sehr exakte Methode, um die Zuverlässigkeit eines Ergebnisses zu prüfen, ist die Kontrolle an der STUDENTschen Verteilung (t-Tabelle), auf die der interessierte Leser hiermit verwiesen sei.

LITERATURVERZEICHNIS

H. L. Anspacher, „Bleibendes und Vergängliches aus der Deutschen Wehrmachtspsychologie", Mitteilungen des Berufsverbandes Deutscher Psychologen e. V., Hamburg — 3. Jahrgang, Nr. 11 — 1949

F. Baumgarten, „Die Berufseignungsprüfungen", R. Oldenbourg, München und Berlin — 1928

Boring-Langfeld-Weld, „Foundations of Psychology"

E. B. Greene, „Measurements of Human Behavior", The Odyssey Press, New York — 1941

H. E. Garrett, „Statistics in Psychology and Education", Longmans, Green and Co., New York, London, Toronto — 1950

J. P. Guilford, „Fundamental Statistics in Psychology and Education", McGraw-Hill Book Company, New York und London — 1942

H. Hosemann, „Die Grundlagen der statistischen Methoden für Mediziner und Biologen", Stuttgart — 1949

R. W. Husband, „Applied Psychology", Harper & Brothers, New York — 1949

F. Klezl-Norberg, „Allgemeine Methodenlehre der Statistik", Springer-Verlag, Wien — 1946

R. Meili, „Psychologische Diagnostik", Alfred Meili, Schaffhausen — 1937

J. L. Mursell, „Psychological Testing", Longmans, Green and Co., New York, London, Toronto — 1947

K. Oppler und E. Rosenthal-Pelldram, „Die Neugestaltung des öffentlichen Dienstes", Kommentator-Verlag, Frankfurt/M. — 1950

Th. Scharmann und W. Dörrhöfer, „Berufsbilder aus Verwaltung und Wirtschaft", Kommentator-Verlag, Frankfurt am Main — 1951

J. Tiffin, „Industrial Psychology", Prentice-Hall, New York — 1946

F. Türk und W. Dörrhöfer, „Neuzeitliche Methoden der Personalauslese", Kommentator-Verlag, Frankfurt am Main — 1950

P. E. Vernon, „Psychological Tests in the Royal Navy, Army and A. T. S.", Occupational Psychology, London — April 1947

P. E. Vernon, „The Structure of Practical Abilities", Occupational Psychology, London — 1949

L. V. Webb und M. Shottwell, „Testing in the Elementary School", Farrar & Rinchart, New York — 1939

E. Weber, „Grundriß der biologischen Statistik", Verlag Gust. Fischer, Jena — 1948

K. Wilde, „Die Frage der Sicherheit in der psychologischen Diagnose I", Psychologische Rundschau I/1, Göttingen — 1949

Schriften
der Deutschen Gesellschaft für Personalwesen e.V.

Nr. 1 **Der öffentliche Dienst in den Vereinigten Staaten von Amerika**
Ein Reisebericht, 192 Seiten, Halbleinen . . . . . DM **7.50**
(Veröffentlicht als Bd. 3 in der wissenschaftlichen Schriftenreihe des Institutes zur Förderung öffentlicher Angelegenheiten e. V.)

Nr. 2 **Neuzeitliche Methoden der Personalauslese**
von Fritz Türk und Walter Dörrhöfer
68 Seiten, kartoniert mit Schutzumschlag . . . . . DM **3.60**
(Veröffentlicht als Bd. 4 in der wissenschaftlichen Schriftenreihe des Institutes zur Förderung öffentlicher Angelegenheiten e. V.)
Zweite (unveränderte) Auflage

Nr. 3 **Die Neugestaltung des öffentlichen Dienstes - Grundlagen und Probleme**
von Kurt Oppler und Erich Rosenthal-Pelldram
80 Seiten, kartoniert mit Schutzumschlag . . . . . DM **2.90**
(Veröffentlicht als Bd. 5 in der wissenschaftlichen Schriftenreihe des Institutes zur Förderung öffentlicher Angelegenheiten e. V.)

Nr. 4 **Personalblatt** Sammelband 1948 - 1950
290 Seiten, Halbleinen . . . . . DM **7.50**

Nr. 5 **Behörde und Publikum**
24 Seiten, kartoniert mit Schutzumschlag . . . . . DM **-.25**

Nr. 6 **Berufsbilder aus Verwaltung und Wirtschaft**
von Theodor Scharmann und Walter Dörrhöfer
58 Seiten, kartoniert mit Schutzumschlag . . . . . DM **3.60**

Nr. 7 **Der Test in der Eignungsuntersuchung**
von Anneliese Kühneck
Eine Darstellung der gebräuchlichen Methoden bei der Ausarbeitung und Bewertung schriftlicher Testverfahren.
68 Seiten, kartoniert mit Schutzumschlag . . . . . DM **3.60**

Nr. 8 **Probleme des öffentlichen Dienstes in England, Frankreich und den Vereinigten Staaten von Amerika**
(Tagungsbericht)
58 Seiten, kartoniert . . . . . DM **1.50**

In Vorbereitung:

Nr. 9 **Bürger und Behörden**
32 Seiten, kartoniert mit Schutzumschlag . . . . . DM **-.60**