Diether Hopf

Möglichkeiten und Grenzen der Anwendung von Tests

first published in:
Funk-Kolleg pädagogische Psychologie, Bd. 2: Lernen und Instruktion / hrsg. v. M. Hofer ... - Frankfurt (Main) : Fischer, 1973, S. 302-312

Postprint published at the Institutional Repository of Potsdam University:
In: Postprints der Universität Potsdam
Humanwissenschaftliche Reihe ; 99
http://opus.kobv.de/ubp/volltexte/2009/3636/
http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-36367

Postprints der Universität Potsdam
Humanwissenschaftliche Reihe ; 99

DIETHER HOPF

Möglichkeiten und Grenzen der Anwendung von Tests*

1. Eigenschaften psychologischer Tests

Beim täglichen Umgang mit anderen Menschen, sei es im Beruf, im Bekanntenkreis oder bei zufälligen Begegnungen, fällen wir fortwährend Urteile über Eigenschaften, Fähigkeiten und Motive anderer Personen. Im allgemeinen scheinen die Diagnosen für den täglichen Bedarf hinreichend genau zu sein; auch lassen sich Irrtümer gewöhnlich schnell berichtigen. Erst wenn auf der Grundlage unseres Urteils Entscheidungen gefällt werden, die nicht mehr ohne weiteres revidierbar sind, wird die Unsicherheit der Diagnosen deutlich, die auf bloßer Menschenkenntnis, Gesprächen oder den üblichen Prüfungen beruhen.

Hier können Tests Verbesserungen herbeiführen, da bei ihnen aufgrund einer wissenschaftlichen Standardisierung die Subjektivität des Beurteilers und die Einwirkung des Prüfers auf den Prüfling in der Beurteilungssituation eingeschränkt sind. Durch Tests kann man mit größerer Sicherheit Aussagen darüber machen, in welchem Grade sich Menschen in bezug auf bestimmte psychische Merkmale, z. B. Intelligenz oder Aggressivität, voneinander unterscheiden.

Tests haben die Aufgabe, von einem begrenzten Verhaltensausschnitt über etwas, was nicht in ganzem Umfang zutage tritt oder was sich erst in Zukunft ereignen wird, Voraussagen zu machen. Das im Test geprüfte Verhalten wird als »Symptom« für eine Eigenschaft oder Fähigkeit verstanden, auf die es verweist; der Zusammenhang von Symptom und der zu messenden Eigenschaft oder Fähigkeit, d. h. dem Kriterium, muß nachgewiesen sein. Beispielsweise kann ein muttersprachlicher Wortschatztest, der aus 40 Wörtern besteht, erst dann als brauchbar gelten, wenn sich zeigen läßt, daß zwischen der Leistung im Test und der gesamten Wortkenntnis der Probanden eine enge Beziehung besteht. Dafür ist die Voraussetzung, daß

- das Kriterium (hier: gesamter Wortschatz) genau erforscht und definiert ist;
- der Testinhalt eine repräsentative Stichprobe aus dem Kriteriumsverhalten darstellt;
- der Zusammenhang zwischen Test und Kriterium an einer repräsentativen Stichprobe aus der Bezugsgruppe nachgewiesen wurde.

* Verfaßt im Auftrag der Stiftung Volkswagenwerk. Hannover 1970.

Test und vorausgesagtes Verhalten brauchen allerdings nicht unbedingt einander ähnlich zu sein; wenn der Nachweis eines empirischen Zusammenhangs geführt werden kann, ist der Test verwendbar. So gibt es z. B. aus Bildern bestehende Tests, zu denen sich die Probanden frei äußern. Aus der Analyse der Äußerungen lassen sich Rückschlüsse auf Leistungsstreben, Aggressivität, soziales Kontaktbedürfnis u. a. ziehen. Testinhalt, Testsituation und Testergebnis sind in diesem Fall von dem diagnostizierten Verhalten ganz verschieden. Selbstverständlich muß hierbei nach einer psychologischen Erklärung der empirisch gefundenen Beziehung gesucht werden.

Objektivität

Im Test wird eine Verhaltensstichprobe unter genau bestimmten Umständen beobachtet und registriert. Es ist von entscheidender Bedeutung, daß dabei die Testsituation für alle Probanden möglichst gleich gehalten wird; die individuellen Unterschiede sollen nur durch die Verschiedenheiten zwischen Personen, nicht durch die Verschiedenheiten zwischen Situationen zustandekommen. Zur Standardisierung der Testsituation dienen vor allem folgende Maßnahmen:

- die vorgeschriebene Instruktion, durch die beispielsweise die Motivation beeinflußt wird; für ein und denselben Test darf nicht bei einer Person höchste Anstrengung, bei einer anderen Gleichgültigkeit oder eine spielerische Haltung hervorgerufen werden; für Tests, die *maximale* Leistungen erfassen (z. B. Fähigkeits- und Schulleistungstests) muß eine andere Motivationsstärke induziert werden als für Tests, die *typische* Reaktionen registrieren (z. B. Persönlichkeitstests);
- die Standardisierung des Testmaterials;
- einheitliche Zeitgrenzen;
- Vorschriften für das Verhalten des Testers bei Fragen des Probanden zum Test und Vermeidung von Störungen und Ablenkungen.

Wie alle wissenschaftlichen Beobachtungen wird ein Test also unter kontrollierten Bedingungen durchgeführt. Darüber hinaus lassen sich Testresultate deshalb als objektiv bezeichnen, weil sie unabhängig vom subjektiven Urteil des Testers gewonnen werden; wer immer das Testverhalten eines Probanden auswertet, klassifiziert und interpretiert, muß zum gleichen Ergebnis kommen.

Normen

Voraussetzungen für die Interpretation von Testresultaten sind empirisch gewonnene Normen. Wenn z. B. ein Proband von 40 Wortschatzaufgaben 29 richtig beantwortet, ist dies Ergebnis erst dann verständlich, wenn man weiß, wie sich die Anzahl richtig gelöster Aufgaben in einer repräsentativen Bezugsgruppe verteilt. Normen werden aus solchen Informationen gewonnen.
Durch Normen wird es möglich, einem individuellen Testresultat einen bestimmten Rangplatz zuzuweisen und seine relative Stellung zur Bezugsgruppe zu erkennen. Je genauer die Merkmale des Probanden mit den Merkmalen der Bezugsgruppe übereinstimmen, desto wertvollere Informationen lassen sich aus den Normen gewinnen. Nach Alter, Geschlecht, Vorkenntnissen, Herkunft usw. differenzierte Normentabellen können deshalb den Wert eines Tests beträchtlich erhöhen.

Reliabilität oder Zuverlässigkeit

Es gibt mehrere Verfahren, die Reliabilität eines Tests festzustellen, z. B. die kurzfristige Wiederholung einer Messung an denselben Personen. Ein Test gilt als reliabel, wenn die Resultate relativ stabil sind. Beispielsweise könnte ein Intelligenztest, in dem ein Proband einen Intelligenzquotienten (IQ) von 120 erhält, nicht als zuverlässig gelten, wenn dieselbe Person bei einer Wiederholungsmessung nach kurzem Abstand einen IQ von 90 erreichte. Andererseits besitzen jedoch auch sehr sorgfältig konstruierte und empirisch überprüfte Tests keine vollkommene Reliabilität; die Testresultate schwanken bei Wiederholungsmessungen in einem bestimmten Spielraum. Gründe hierfür sind z. B. Veränderungen der jeweiligen äußeren Bedingungen (Wetter, Lärm etc.) oder der Disposition des Probanden (Müdigkeit, Krankheit usw.). Die je nach Reliabilität variierende Schwankungsbreite wird durch den *Meßfehler* bezeichnet. So beträgt beispielsweise bei den gebräuchlichen Intelligenztests ein Meßfehler ungefähr $\pm$ 5 IQ-Punkte. Daraus folgt, daß es unzulässig wäre, einen Probanden mit dem gemessenen IQ von 116 für intelligenter zu halten als einen mit dem IQ von 110: Man muß vielmehr annehmen, daß die Intelligenzquotienten der Probanden mit einer Wahrscheinlichkeit von 2:1 bei dem einen zwischen 121 und 111 (116 $\pm$ 5), bei dem anderen zwischen 115 und 105 (110 $\pm$ 5) liegen; bei einer Wiederholung des Tests könnte sich die Rangfolge der

Probanden umkehren. Aus diesem Grunde geben moderne Tests das Ergebnis nicht mehr in Form eines Punktwertes an, sondern als »Band«, innerhalb dessen der gesuchte Wert liegt. Solange sich zwei Bänder bei einem Vergleich überlappen, gilt die Differenz als nicht interpretierbar.
Die zur Illustration des Meßfehlers als Beispiel gewählten Intelligenztests gehören zu den zuverlässigsten Testverfahren. Eine höhere Präzision ist bei psychologischen Messungen kaum zu erreichen. Man muß sich klar machen, daß der Meßfehler anderer Beurteilungsverfahren in der Regel sehr viel größer ist und nur aufgrund fehlender empirischer Untersuchungen unbekannt bleibt. Bei einem guten Schulleistungstest, dessen Skala sich von 60 bis 140, also über 80 Punkte, erstreckt, würde der dem obengegebenen Beispiel entsprechende Meßfehler ebenfalls ca. ± 5 Punkte betragen; bei Schulzensuren dagegen müßte man bei einer Skala von nur 6 Punkten (1 bis 6) mit einem Meßfehler von etwa ± 1 Zensur rechnen.

Validität oder Gültigkeit

Ein Test ist in dem Grade valide, in dem er mißt, was er messen soll. Die Bestimmung der Validität erfordert in der Regel eine empirische Überprüfung des Zusammenhangs des Tests mit einem unabhängigen Kriterium. Bei einem Test, der beispielsweise aus einer großen Bewerberzahl für den Beruf des Flugzeugpiloten diejenigen herausfinden soll, die mit größter Wahrscheinlichkeit die Ausbildung erfolgreich durchlaufen und den Anforderungen des Berufs genügen werden, muß eine hinreichende Korrelation zwischen den Testergebnissen und dem Kriterium — Ausbildungs- bzw. Berufserfolg — bestehen. Nur wenn sich ein enger Zusammenhang hat nachweisen lassen, kann der Test als Ausleseinstrument künftigen Bewerbern vorgelegt werden. Die Wahl des Ausbildungserfolges, nicht jedoch der Berufsleistung, als Validierungskriterium wäre in diesem Beispiel allerdings problematisch, solange über die Korrelation zwischen Ausbildungsergebnis und Berufserfolg nichts bekannt ist; selbst eine hohe Korrelation zwischen Testleistung und Ausbildungserfolg könnte darüber hinwegtäuschen, daß unter Umständen gerade die geeigneten Bewerber zurückgewiesen wurden.
Die Korrelationen zwischen Test und Kriterium sind in aller Regel nicht besonders hoch. Man muß deshalb bei der Diagnose und Prognose mit einem *Schätzfehler* rechnen, dessen Größe sich angeben läßt, wenn Validitätsuntersuchungen durchgeführt worden sind. Aus dem Schätzfehler

läßt sich ableiten, wie oft aufgrund der Testresultate unvermeidliche Fehlentscheidungen getroffen werden, in unserem Beispiel: wieviele ungeeignete Bewerber aufgenommen und wieviele geeignete abgewiesen werden.
Tests sollten zur allgemeinen Verwendung nur dann freigegeben werden, wenn ihre Reliabilität und Validität durch empirische Untersuchungen gesichert worden sind. Auch dann spielen jedoch bei allen psychologischen Messungen Meß- und Schätzfehler eine große Rolle. Sie sind besonders zu beachten, wenn Entscheidungen auf Testresultate gegründet werden sollen.

2. Anwendung psychologischer Tests

Tests lassen sich in einer Vielzahl von Bereichen wie z. B. Schule, Universität, Klinik, Betrieb, Beruf, Erziehungsberatung, Militär und Verkehr zum Zwecke der Auslese, Klassifikation und Forschung verwenden.

Auslese

Die am weitesten verbreitete und folgenreichste Verwendungsart von Tests dürfte derzeit darin bestehen, daß mit ihrer Hilfe Ausleseentscheidungen getroffen werden. So werden beispielsweise Tests häufig benutzt beim Übergang von der Grundschule zum Gymnasium; zur Bestimmung der Schulreife; bei der Personalauslese für Behörden und Betriebe; zur Identifizierung psychischer Krankheiten etc. Durch Ausleseentscheidungen werden zwei Gruppen geschaffen: die Ausgewählten und die Zurückgewiesenen. In der Regel wird die Güte der Ausleseentscheidung nur am Erfolg oder Mißerfolg bei den Ausgewählten gemessen; die Fehlentscheidungen hinsichtlich der Zurückgewiesenen geraten aus dem Blickfeld.
Die Auslesesituation ist dadurch charakterisiert, daß diejenigen Personen, die ein bestimmtes Merkmal in einer bestimmten Ausprägung besitzen, z. B. hohe Intelligenz, ausfindig gemacht werden sollen. Bei diesem Vorgang sind aufgrund der unvollkommenen Reliabilität und Validität praktisch aller psychologischen Beurteilungsverfahren Fehler unvermeidlich.

Die bei der Selektion entstehenden Fehler lassen sich in zwei Typen unterteilen: A) ungerechtfertigte positive Auslese von Probanden, die fälschlicherweise als Merkmalsträger diagnostiziert werden; B) ungerechtfertigte negative Auslese von Probanden, die zurückgewiesen werden, obwohl sie Merkmalsträger sind. Am Beispiel der Übergangsauslese für das Gymnasium heißt dies, daß einerseits eine Reihe von Kindern aufgenommen wird, die nach den Normen des derzeitigen Gymnasiums nicht aufgenommen werden dürfte (Fehlertyp A) und daß andererseits potentielle Gymnasiasten zu Unrecht abgewiesen werden (Fehlertyp B).

Vorausgesetzt, es besteht eine hinreichende Korrelation zwischen Test und Kriterium, lassen sich die beiden Arten der Fehler mit Hilfe unterschiedlicher Strategien verringern: möchte man den Fehlertyp A minimieren, muß man die Aufnahmequote drastisch senken (in unserem Beispiel: von ca. 25% auf 5%); will man Fehler des Typs B vermeiden, muß man die Aufnahmequote erheblich vergrößern (z. B. von 25% auf 70%). Mit der Verringerung des einen Fehlertyps ist notwendigerweise eine Vergrößerung des anderen verbunden.

Wenn Tests zur Selektion verwandt werden, muß eine Vorentscheidung getroffen werden, welchen Fehlertyp man vermeiden möchte. Bei der Auslese von Piloten für die zivile Luftfahrt dürfte es vor allem darauf ankommen, Fehler vom Typ A auszuschließen, da das Versagen eines Flugzeugführers verheerende Folgen haben kann. Es ließe sich begründen, warum die Zurückweisung auch geeigneter Bewerber, so hart sie für den einzelnen sein mag, in Kauf genommen werden muß; die »institutionelle« Entscheidung hat vor der »individuellen« den Vorrang. Bei der Auswahl psychiatrischer Patienten, die sich einem irreversiblen Eingriff unterziehen sollen, muß ebenfalls alles dafür getan werden, daß Fehler vom Typ A nicht unterlaufen.

Bei der Auslese fürs Gymnasium oder bei einem Numerus clausus für die Hochschule liegt jedoch der Fall aus offensichtlichen Gründen anders. Institutionelle Entscheidungen besitzen hier nicht dieselbe Legitimationsbasis; folglich muß alles getan werden, um die Zurückweisung Geeigneter zu vermeiden. Um ein Bild von der Größenordnung der Fehler zu geben, mit denen in diesen Bereichen gerechnet werden muß: Bei der Validität der schulischen Aufnahmeprüfungen dürfte die Summe der Fehlentscheidungen (Fehler vom Typ A und B) mindestens 25% betragen; besäße man in Deutschland Tests, deren Validität so hoch wäre wie die Validität

der besten in den USA verwendeten College-Eingangstests, so würden im Falle eines auf zwei Drittel der Bewerber beschränkten Numerus clausus im ganzen knapp 40 % Fehlentscheidungen getroffen. Wenn man sich klar macht, welche individuellen Veränderungen und Entwicklungsschwankungen in den Zeiträumen, auf die sich die Prognosen beziehen, eher die Regel als die Ausnahme darstellen, wird die hohe Fehlerzahl freilich verständlich.
Die Nützlichkeit eines Tests bestimmt sich aus der zusätzlichen Information, die er liefert. Beim Beispiel der Schülerauslese bemißt sie sich daran, wie groß der Gewinn an Präzision bei Tests gegenüber der herkömmlichen Aufnahmeprüfung ist. Trotz der relativ hohen Fehlerquoten ist die prognostische Validität guter Tests in der Regel höher als die der Lehrerurteile — wie empirische Untersuchungen gezeigt haben. Allerdings sollte das Instrument »Test« immer nur als zusätzliche Information verwendet werden, da die Kombination mehrerer Verfahren, z. B. Test und Lehrerurteil, meist die Sicherheit der Diagnose erhöht. In England, wo die Übergangsauslese sehr verfeinert wurde und hochvalide Tests gebräuchlich sind, beträgt die Summe der Fehlentscheidungen im Rahmen des bestehenden Systems im besten Fall ca. 15 %. Das bedeutet, gemessen an den ca. 25 % Fehlern, die auf der Grundlage von Schulnoten und -empfehlungen auftreten, zwar eine erhebliche Verbesserung, jedoch noch immer ein so hohes Maß an Ungenauigkeit, daß als Konsequenz die Beseitigung des selektiven Schulsystems eingeleitet wurde.

Kriteriumsverhalten und Selektionsstrategie

Bei der Verwendung von Tests zur Auslese ist die richtige Wahl und die genaue Kenntnis des Kriteriums notwendige Voraussetzung für die Bestimmung der besten Selektionsstrategie. Das gilt insbesondere für die zahlreichen Fälle, in denen mehr als ein Testresultat die Entscheidungsbasis bildet. Bei einer Zulassungsprüfung zum Medizinstudium z. B. könnten Untersuchungen über den Zusammenhang von Schulnoten und Studienleistungen möglicherweise nahelegen, schlechte Schulleistungen in Physik durch gute Biologiekenntnisse als ausgeglichen zu betrachten; bei der Auswahl von Busfahrern wäre es dagegen sicher nicht angezeigt, sehr niedrige Reaktionsgeschwindigkeit oder Farbenblindheit durch große Pünktlichkeit aufzuwiegen. Vor der Testkonstruktion und vor der Entwicklung einer Strategie, Testresultate zu kombinieren und zu gewichten, ist deshalb eine genaue Er-

forschung des Kriteriumsverhaltens notwendig. Ohne solche Vorkenntnisse ist es nicht möglich, die Folgen einer bestimmten Auslesepraxis zu übersehen. Die Testforschung in Deutschland steht hier noch ganz am Anfang. Um nur ein Beispiel zu nennen: Es gibt so gut wie keine brauchbaren Untersuchungen, die den Zusammenhang von Schul- oder Hochschulerfolg und Berufsleistung überprüft haben. Selbst wenn ein Test in befriedigender Weise das Kriterium Studienerfolg vorhersagte, wüßte niemand, ob nicht unter dem Gesichtspunkt der späteren Berufsleistung gerade die Geeigneten dem Numerus clausus zum Opfer gefallen sind. Ist es z. B. so sicher, daß die hervorragenden Lateinstudenten auch die besten Lateinlehrer sein werden? Für die Konstruktion von Hochschuleingangstests müßten dementsprechend die generellen Bedingungen des Studien- *und* Berufserfolges ermittelt werden, auf die hin dann die Tests zu validieren wären.

Unerwünschte Stabilisierungseffekte

Voraussetzung für die Verwendung von Tests zur Auslese ist also die Untersuchung ihrer prognostischen Validität. Ein Test ist jedoch nur so lange valide, wie das Kriterium unverändert bleibt. Jede Veränderung der Bedingungen, unter denen das Kriteriumsverhalten ursprünglich bestimmt wurde, machen eine erneute Validierung notwendig. Eine der großen Gefahren bei der Verwendung von Tests besteht darin, daß gewöhnlich über lange Jahre weder die Normen berichtigt noch die Validität überprüft werden, selbst wenn, wie derzeit im Bildungswesen, rasche und massive Veränderungen zu erkennen sind. Tests, die nicht den neuen Bedingungen angeglichen werden, wählen nach den Normen des vergangenen Systems aus und führen dadurch entweder zum Mißerfolg der Ausgewählten oder verhindern die Durchsetzung von Reformen.

Neben den bei einer Selektion in der Regel unvermeidlichen Fehlern muß eine weitere Gefahr der Testverwendung erwähnt werden, die insbesondere im Bereich des Bildungswesens besteht. Wenn Tests Verteilerfunktion erhalten, wirken sie auf die Zubringerinstitutionen zurück. So hat in England die sehr weitgehende Verwendung von Tests bei der Übergangsauslese zur Sekundarstufe massive Effekte auf den Lehrplan und die Unterrichtsformen der Primarschule gehabt, in der die Schüler teilweise jahrelang einseitig auf die Prüfung vorbereitet wurden, so daß kaum Möglichkeiten bestanden, innovative Unterrichtsformen und -inhalte zu erproben. Nur wenn Tests ständig neu auf die Ziele des Schul-

oder Hochschulunterrichts abgestimmt werden — und zwar dergestalt, daß die Tests den Veränderungen des Unterrichts und der Weiterentwicklung des Curriculum folgen, nicht umgekehrt — läßt sich diese Gefahr bannen. Das erfordert jedoch ständige, auf Forschung beruhende Revision.

Klassifikation

Die geschilderten Gefahren bei der Verwendung von Tests zu Auslesezwecken — wobei immer zu beachten ist, daß die Anwendung von validen Tests trotz des hohen Fehlerspielraums den gängigen individuellen Urteilen vorzuziehen ist — treten in milderer Form auf, wenn Tests der Klassifikation dienen. Hierbei wird kein Proband aufgrund der Diagnose zurückgewiesen, sondern jeder einzelne erfährt die ihm entsprechende Behandlung. Klassifikationsentscheidungen werden z. B. in der Berufsberatung getroffen, indem dem Ratsuchenden die Erfolgschancen von Probanden mit ähnlichen Testwerten mitgeteilt werden. Auf ihrer Grundlage kann er selbst eine Entscheidung treffen, ob er sich der Prognose entsprechend verhält oder das Risiko einer abweichenden Entscheidung eingeht; beim Militär, wenn es darum geht, jeden an den Platz zu stellen, für den er am besten geeignet ist; im klinischen Bereich, wo für jeden Patienten die optimale Therapie gefunden werden muß; in der Gesamtschule, in der die Schüler nach Interesse und Leistung für bestimmte Gruppen optieren.
Hier wird deutlich, welche Funktion eine Beratung anhand von Testergebnissen haben kann: die Wahrscheinlichkeit von Erfolg oder Mißerfolg, von zu erwartenden Widerständen etc. wird mitgeteilt und somit die individuelle Entscheidung auf eine bessere Grundlage gestellt.
Der Vorzug einer Testverwendung zur Klassifikation besteht zunächst darin, daß die mit Hilfe der Tests gewonnenen Informationen besser genutzt werden als bei der Selektion, da es keine Zurückgewiesenen gibt, deren Daten verlorengehen. Zum anderen ist dabei die Möglichkeit einer nachträglichen Revision der auch hier unvermeidlichen Fehlentscheidungen meist wesentlich größer, so daß weitere Informationen über die Richtigkeit der anfänglichen Verteilung sämtlicher Probanden gesammelt werden können. Im Gegensatz dazu ist bei der Selektion die Ablehnung eines Probanden in der Regel eine endgültige Entscheidung; nur bei den Ausgewählten besteht die Möglichkeit, die Entscheidung noch zu korrigieren. Bei der Selektion unterzieht sich zudem meist nur eine kleine Gruppe (z. B. Bewerber) dem

Test; Geeignete, die sich aus irgendwelchen Gründen nicht bewerben — z. B. um Aufnahme ins Gymnasium — bleiben unerkannt. Dieser schwerwiegende Nachteil kann durch die Diagnose, Beratung und Klassifikation aller Betroffenen vermieden werden.

Forschung

Die Verwendung von Tests in der Forschung soll hier nicht ausführlich erörtert werden. Daß für die Zwecke der Auslese und Klassifikation umfangreiche, fortlaufende Forschungsarbeiten erforderlich sind, dürfte aus dem Gesagten deutlich geworden sein. Es versteht sich von selbst, daß zahlreiche Probleme der Psychologie, Soziologie, Erziehungswissenschaft, Psychiatrie usw. ohne Tests nicht in Angriff genommen werden können.

3. Aufwand für die Testkonstruktion

Die Konstruktion guter Tests erfordert für die Herstellung des Testmaterials, die Gewinnung von Normen und die notwendigen Untersuchungen zur Reliabilität und Validität große Investitionen an Zeit, qualifizierter Arbeitskraft und finanziellen Mitteln. So dürfte beispielsweise die Herstellung eines guten Schulleistungstests für eine Klassenstufe in einem Unterrichtsfach derzeit die Arbeitskraft von mindestens drei Experten über einen Zeitraum von ca. $1^1/_2$ Jahren mit Kosten von insgesamt ca. DM 300 000,— erfordern. Diese anhand eines konkreten Falles gewonnene Schätzung gilt freilich nur unter den derzeit schwierigen Bedingungen in Deutschland; sobald ausgebildete Fachleute zur Verfügung stehen und eine gut funktionierende Organisation zur Durchführung der notwendigen empirischen Erprobungen vorhanden ist, dürften die Kosten und der Zeitaufwand erheblich absinken. Der Zeitbedarf für die Bestimmung der prognostischen Validität eines Tests beträgt meist mehrere Jahre. Ihre ständige Überprüfung ist unerläßlich.

Die statistische Aufbereitung und Verrechnung von Daten sowie Service-Funktionen für Testbenutzer lassen sich weitgehend durch Computer wahrnehmen; sie erfordern zwar nur einen geringen Zeitaufwand, jedoch gut ausgebildetes Personal, da sonst die notwendige Forschungsarbeit nicht geleistet werden kann.

Den Investitionskosten stehen die Vorteile gegenüber, die

die Verwendung von Tests mit sich bringen. Diese lassen sich allerdings nur in seltenen Fällen, z. B. bei der Personalauslese in Betrieben, in Form von Kostenersparnis ausdrücken. Tests ermöglichen sparsame Messungen eines vorhandenen Merkmals und ökonomische Prognosen zukünftigen Verhaltens. Mit der nötigen Vorsicht angewandt, stellen sie unschätzbare Entscheidungshilfen für den einzelnen und die Gesellschaft bereit.

4. Zusammenfassung

Zuverlässige und gültige Tests liefern auf ökonomische Weise wertvolle Informationen zur Unterscheidung von Individuen. Sie stellen relativ verläßliche Grundlagen für einen Vergleich bereit und ermöglichen Prognosen. Auslese- und Klassifikationsentscheidungen lassen sich mit Hilfe von Tests sicherer fällen als ohne sie. Tests sind zudem unentbehrliche Hilfsmittel der sozial- und verhaltenswissenschaftlichen Forschung.

Gerade weil Tests in einer komplexen und interdependenten Gesellschaft unerläßliche Entscheidungshilfen darstellen, müssen ihre Grenzen deutlich gemacht und ihre Gefahren aufgewiesen werden. Diese bestehen zum einen darin, daß im Falle der Selektion zahlreiche irreversible Fehlentscheidungen gefällt werden, und zwar selbst dann, wenn sorgfältig konstruierte und erprobte Instrumente die Urteilsgrundlage bilden. Aus diesem Grund wäre es nicht selten wünschenswert, auf eine Selektion zu verzichten und den institutionellen Kontext zu ändern. Anstelle einer punktuellen, meist irreversiblen Auslese wären mit Hilfe von Tests auf Information und Beratung gegründete Entscheidungen zu setzen.

Zum anderen können Tests dadurch unerwünschte Konsequenzen haben, daß sie auf ihren Wirkungsbereich stereotypisierende Effekte ausüben. Das wird immer dann der Fall sein, wenn nicht für ihre ständige Revision und Abstimmung auf die gegebenen Zielvorstellungen gesorgt wird. Nur eine ganz enge Kooperation von Forschung und Testpraxis und die Offenlegung der in die Testkonstruktion und die Testanwendung eingehenden Entscheidungen können deshalb gewährleisten, daß schwerwiegende Folgeschäden vermieden werden.