Kann man ein Auswahlverfahren üben?

Uwe Füllgrabe, Diplom – Psychologe, Psychologieoberrat
Bildungsinstitut der Polizei Niedersachsen, Hann. Münden

1. Kann man Tests trainieren?

Tests spielen bei Auswahlverfahren eine große Rolle. Dies ist verständlich, weil eine gute Personalauswahl dafür sorgt, dass die Zahl der Geeigneten bei den eingestellten Personen hoch und die Zahl der Ungeeigneten gering ist. So sank bei einer deutschen Polizeischule nach der Einführung eines wissenschaftlich fundierten psychologischen Auswahlverfahrens die Zahl der Anwärter, die den Grundlehrgang nicht bestanden auf unter 5% (in einige Jahren auf 1%). Dabei muss darauf hingewiesen werden, dass bei den Personen, die den Lehrgang nicht bestanden, spezifische Motivationsprobleme eine Rolle spielten, etwa das Syndrom des *passiven Lebensstils* (s. Füllgrabe, 2002).

Die Bedeutung der Tests für das erfolgreiche Bestehen eines Auswahlverfahrens hat auch einen bestimmten Trainingsmarkt geschaffen, bestehend aus Büchern für die Bearbeitung von Tests oder sogar konkrete Seminare für Testtraining.

Die Frage ist aber, ob man die Tests oder sogar ein ganzes Auswahlverfahren trainieren kann. Diese Frage kann man nur beantworten, wenn man einmal Personen den gleichen Test ein zweites Mal machen lässt. Dazu gibt es eine Reihe von Erkenntnissen.

Ganz anders sieht es aber aus, wenn man die Frage stellt, in wieweit man sich verbessern würde, wenn man ein ganzes Auswahlverfahren noch einmal durchlaufen könnte. Dies ist durchaus von wissenschaftlichem und auch praktischem Interesse, wird aber angesichts des hohen zeitlichen und finanziellen Aufwandes und angesichts der Tatsache, dass man ja geeignete Bewerber gefunden hat, wohl kaum untersucht. Deshalb ist mir persönlich, außer der in diesem Artikel erwähnten Studie, keine weitere Untersuchung zu dieser Problematik bekannt.

Die Seltenheit dieser Daten ist auch der Grund, warum in diesem Artikel die Frage nach der Trainierbarkeit von Auswahlverfahren beantwortet wird.

2. Probleme der Intelligenzmessung

Wie die Intelligenzforschung zeigt, ist Intelligenz keine starre, unveränderliche Größe. Bezeichnenderweise hat Binet, der Vater der Intelligenztests, seinen Intelligenztest dazu benutzt, unterdurchschnittlich intelligente Kinder zu trainieren. Man kann deshalb Intelligenz auch als *Fähigkeit* oder, wie Horn, als das Ergebnis einer *Leistung* betrachten, wobei sich für Menschen individuelle Unterschiede zeigen können. Genauso wie man sportliche oder andere Fähigkeiten erwerben oder verbessern kann, kann man unter bestimmten Voraussetzungen seine Intelligenz*leistung* steigern. Dies bedeutet nicht, dass man Intelligenz einfach mechanisch einüben kann; vielmehr kann man seine Leistungen bei Intelligenztests nur dadurch steigern, dass man bessere geistige Problemlösungsstrategien entwickelt, z.B. durch Abbau eines impulsiven Bearbeitungsstils (s. Füllgrabe, 1978). Es wird also verständlich, warum beim erneuten Bearbeiten des gleichen Tests höhere Leistungen erzielt werden können.

Da aber die (geistigen) Probleme, die in verschiedenen Situationen, bei verschiedenen Berufszweigen u.ä. höchst unterschiedlich sein können, ist verständlich, dass es keine einzige globale Intelligenz gibt. Man muss sich vielmehr die Intelligenz eines Menschen

als aus vielen verschiedenen Fähigkeiten zusammengesetzt vorstellen. Dies führt dazu, dass sich zwei Menschen bezüglich der einzelnen Fähigkeiten extrem voneinander unterscheiden können, obwohl sie rein rechnerisch die gleiche Intelligenz (den gleichen „IQ“) aufweisen. Die praktische Konsequenz daraus ist: Trotz gleich hoher Intelligenz eignet sich die eine Person für

einen bestimmten Beruf, die andere aber nicht, weil sie nicht die für diesen Beruf notwendigen Fähigkeiten besitzt. Tatsächlich wurde bei der Eichstichprobe des Auswahlverfahrens einer deutschen Polizeischule (1972) gefunden, dass gute Lehrgangsteilnehmer und solche, die den Lehrgang nicht bestanden hatten, sich nicht im Gesamt -c- Wert (den Ausdruck der gesamten intellektuellen Leistungsfähigkeit) unterschieden, sondern nur bezüglich der einzelnen Untertests. Dies ist z.B. der Grund dafür, dass der Gesamt -c- Wert nicht als Auswahlkriterien herangezogen werden sollte!

Die Spezifität intellektueller Eigenschaften wirft ein praktisches Problem auf: Da man derartige Fähigkeiten immer weiter untergliedern kann (z.B. fand man **zwei** Faktoren des räumlichen Vorstellungsvermögens), stellen die Benennungen einzelner Intelligenztests nur allgemeine Beschreibungen dar. Es ist deshalb durchaus möglich, dass zwei Tests, die den gleichen Namen tragen, voneinander abweichende Ergebnisse liefern **können**. Dazu ein Beispiel aus Greif (1972):

Wie wenig „logische“ und formale Ähnlichkeitsannahmen über die empirische Vergleichbarkeit aussagen, zeigt ein Vergleich des WIT- Untertests „Wortauswahl“ (WA) mit dem IST- Test (von Amthauer) „Gleiche Wortbedeutung“ (GW). Man könnte meinen, dass beide Tests als Paralleltests anzusehen wären. In beiden wird von der untersuchten Person im Prinzip ähnliches verlangt. Nur die „Vorzeichen“ ändern sich dabei. Beim Subtest GW muss zu einem vorgegebenen Wort ein Wort mit gleicher Bedeutung und beim Subtest WA ein Wort mit verschiedener Bedeutung gesucht werden. Beide Subtests korrelieren jedoch nur mit $r = .28$. Dies zeigt, dass eine einfache, „logische“ Ähnlichkeitsbeurteilung eine *empirische* Ähnlichkeitsuntersuchung nicht ersetzen kann.

Ein weiteres Beispiel dafür, dass man empirisch vorgehen muss und nicht etwa Kriterien „am grünen Tisch“ aufstellen darf, ergab sich bei dem in diesem Artikel besprochenen Auswahlverfahren bei der Erfassung des Arbeitsverhaltens. Auf S. 4 des LPS von Horn findet man eine Variante des Pauli – Tests, von dem bei der Eichstichprobe 4 Spalten bearbeitet wurden.

Man könnte nun z.B. erwarten, dass folgende Werte die Anstrengungsbereitschaft anzeigen könnten:

- „Die Streuung“ = Wert der Spalte mit dem höchsten Wert minus Wert der Spalte mit dem niedrigsten Wert.
- „Endspurtphänomen“ = Wert der 4. Spalte minus Wert der 1. Spalte
- „Durchschnittswert“ D = Durchschnittlicher Wert für die in den 4 Spalten bearbeiteten Aufgaben

Tatsächlich zeigte sich später, dass nur der D- Wert, aber nicht die beiden anderen Werte, eine diagnostische Bedeutung hinsichtlich Anstrengungsbereitschaft besaß. Er erwies sich insgesamt auch als bester Vorhersagewert für die Lehrgangsleistung! Der D- Wert zeigt gewissermaßen die Bereitschaft an, am Ende eines anstrengenden *Auswahlverfahrens* noch einmal „alle Kräfte zu mobilisieren“. Das bedeutet aber auch, dass man hier nicht eine allgemeine situationsunabhängige Leistungsmotivation erfasst. Dies erklärt auch die Tatsache, dass auch Bewerber, die später den Lehrgang nicht bestanden, durchaus über dem festgelegten Mindestwert für D gelegen hatten. Auch ergab sich nur eine Punktbiseriale Korrelation von +0,27 für die Hoch- und

Niedrigmotivierten bezüglich Lehrgangsverhalten bzw.-
 motivation und D - Wert. Interviews mit Personen, die den Lehrgang nicht bestanden hatten, gaben eine Erklärung dafür: Wer am
 Anfang den Schwierigkeitsgrad des Lehrganges unterschätzte und / oder unangemessene Lerntechniken benutzte, konnte natürlich
 keine (im Test ermittelte) maximale Motivation entfalten.

3. Was misst das Leistungs-Prüf-System (LPS) von Horn?

Unter Berücksichtigung der obigen Erkenntnisse erscheint es notwendig, die einzelnen Untertests des L-P-S differenziert daraufhin
 zu untersuchen, was sie erfassen:

- Untertests 1 + 2 = Fähigkeiten, die mit Allgemeinbildung und
 Rechtschreibung zusammenhängen.
- Untertests 3 + 4 = Erkennen von Regeln und Gesetzmäßigkeiten („logisches
 Denken“)
- Untertest 3 = visuell orientierte Aufgabentypen, dem „Matrizentest“
 nachempfunden und nach Ansicht von Horn von
 schulischer Vorbildung unabhängig, während
- Untertest 4 = schulische Einflüsse beinhaltet, weil er die Kenntnis von
 Zahlen und Buchstaben voraussetzt.
- Untertest 5 + 6 = „Wortflüssigkeit“, d.h. die Fähigkeit, schnell viele Wörter zu
 produzieren. Horn weist ausdrücklich darauf hin, dass
 hohe Leistungen bei diesen Tests bei einigen
 Berufsgruppen nachteilig sein kann (z.B. „in
 Führungsstellen und bei Arbeiten, die mehr technische
 Fähigkeiten verlangen.“) Horn (1962) schreibt „Spricht
 einer zu schnell, schalten manche einfach ab.“
- Untertest 7 - 10: = räumliche Vorstellungsvermögen, technische Fähigkeiten.
- Untertest 10 = Erkennen von Wesentlichem, geistige Beweglichkeit,
 geistige Unabhängigkeit.

Begabten, aber

auch für „Bürokräfte, von denen das Erraten schlechtgeschriebener Schriftsätze gefordert wird.“

Untertest 13 + 14 = Aufgaben, die Aufmerksamkeit und Konzentration erfordern und das Wahrnehmungstempo erfassen.

Untertest 13 = ist wichtig für das schnelle Erledigen von Routineaufgaben, wobei jemand sich keinen Fehler leisten kann, weil dies eine Kettenwirkung hätte. Horn nennt hier die Berufe: Buchhalter, Statistiker, Berufe in der Atomindustrie.

Untertest 14 = ist stärker als Untertest 13 auf Wahrnehmungsgeschwindigkeit ausgerichtet.

Die Normwerte des LPS von Horn werden in c- Werten – von 1 bis 9 ausgedrückt- wobei der Mittelwert 5 beträgt und 9 den höchsten Wert darstellt. Die c- Werte lassen sich nach einer einfachen Formel auch in IQ- Werte, oder andere Normwerte umrechnen.

4. Veränderung der Werte bei einer Testwiederholung

Wenn man einen Test mit der gleichen Personengruppe wiederholt, können sich die Testwerte durchaus verändern. Dies zeigt Horn (1962) hinsichtlich der Retest – Reliabilität seines LPS (s. Tabelle 1).

(hier etwa Tabelle 1 einfügen)

Aus der Tabelle 1 geht z.B. hervor, dass für den Gesamt – c- Wert die Personen beim 1. Testdurchgang *im Durchschnitt* 254,9 Aufgaben richtig lösten, im 2. Testdurchgang aber 298,8 Aufgaben, also 43, 9 Aufgaben mehr. Dies ist nur für den unverständlich, der meint, Intelligenz sei ein festes, unveränderliches Merkmal eines Menschen. In Wirklichkeit geschieht aber folgendes: Aus den Leistungen bei einem Test wird auf „die Intelligenz“ geschlossen. Die Intelligenzleistung ist dabei das Ergebnis eines *Entscheidungsprozesses* beim Lösen der Aufgaben. Und durch Verbesserung dieser Entscheidungsprozesse kann auch die Intelligenztestleistung verbessert werden (Füllgrabe, 1978).

Dass sich die Anzahl der gelösten Aufgaben im Durchschnitt vergrößert hat, hat aber nicht unbedingt einen großen Einfluss auf die diagnostische Qualität des LPS. Denn die Korrelation von $r = +0,95$ für den Gesamt- c Wert besagt, dass in der Gruppe die *Reihenfolge* der Leistungen gleich geblieben ist. D. h. (vereinfacht gesagt, unter der Voraussetzung, dass die Korrelation perfekt $r = +1,0$ wäre): Wer im 1. Test die beste Leistung erbracht hatte, hat dies auch beim 2. Test, der Zweitbeste war auch beim 2. Test der Zweitbeste, der Schlechteste im 1. Test war auch im 2. Test der Schlechteste usw. Da der IQ und auch die anderen Normwerte für Intelligenz sich aus dem Vergleich der Einzelleistung mit dem Durchschnitt der Gruppe ergeben, also eine Rangreihe angeben, ist entscheidend, dass bei Testwiederholungen die Rangreihe der Leistungen gleich bleibt.

Problematisch ist aber, wenn die Korrelationen bei Testwiederholungen nicht hoch sind. Bei Testwiederholungen korrelieren beispielsweise die Werte für den Untertest 13 nur $+0,38$. Dies bedeutet konkret, dass bezüglich dieses Faktors die „Menschen im allgemeinen weniger gleichmäßig und genau arbeiten, als man sich dieses oft erhofft.“ (Horn, 1962). Am Rande soll noch darauf hingewiesen werden, dass auch einzelne Testwerte nicht unbedingt hoch miteinander korrelieren: Die Fehlerzahl bei Untertest 13 (Testwert für 13-) korreliert nur wenig mit der intellektuellen Leistung, ist für einen *Intelligenztest* weniger bedeutsam.

5. Effekte bei einer Wiederholung des Auswahlverfahrens

Tabelle 2
Veränderungen von Testwerten bei Testwiederholungen

	1. Stichprobe	2. Stichprobe männlich	2. Stichprobe weiblich	3. Stichprobe
LPS: Gesamt C:	+1,08	+0,42	+0,53	+0,82
	-1,0 - + 3,8	-2,4 - +1,6	-0,8 - +1,2	-98 - +3,0
	$p < .01$	$p < .01$	$p < .01$	$p < .01$
Test 1 + 2	+0,51	+0,24	+0,21	+0,33
	-1,0 - +2,0	- 1,0 - +2,0	-1,0 - +2,0	
	$p < .01$	.10 > $p > .05$	n.s.	$p < .01$

Test 3	+0,87 -2,0 - +4,0 p < .01	+0,41 p < .01	+0,57 p < .01	+0,68 p < .01
Test 4	+0,68 -2,0 - +5,0 p < .01	+0,5 -1,0 - +3,0 .05 > p > .01	+0,5 -1,0 - +2,0 n.s.	+0,47 -3,0 - +4,0 p < .01
Test 5	+1,65 -2,0 - +4,0 p < .01	+0,68 -1,0 - +3,0 p < .01	+1,29 0 - +3,0 p < .01	+1,45 -1,0 - +5,0 p < .01
Test 6	+0,73 -2,0 - +3,0 p < .01	+0,15 -2,0 - +2,0 n.s.	+0,36 -1,0 - +2,0 n.s.	+0,51 -2,0 - +3,0 p < .01
Test 7	+0,53 -1,0 - +4,0 p < .01	+0,18 -2,0 - +4,0 n.s.	+0,07 -2,0 - +2,0 n.s.	+0,66 -2,0 - +3,0 p < .01
Test 8	+0,68 -2,0 - +3,0 p < .01	+0,41 -2,0 - +3,0 .05 > p > .01	+0,21 -2,0 - +2,0 n.s.	+0,49 -3,0 - +3,0 p < .01
Test 9	+0,32 -2,0 - +2,0 p < .05	+0,06 -2,0 - +3,0 n.s.	+0,5 -3,0 - +4,0 n.s.	+0,26 -2,0 - +3,0 .05 .01
Test 10	+0,81 -2,0 - +5,0 p < .01	+1,0 -1,0 - +4,0 p < .01	+0,64 0 - +3,0 p < .05	+8,85 -3,0 - +6,0 p < .01
Test 11	+0,87 -2,0 - +3,0 p < .01	+0,35 -3,0 - +3,0 n.s.	+0,71 -1,0 - +2,0 p < .02	+0,29 -2,0 - +4,0 .05 = p.01
Test 12	+0,62 -1,0 - +3,0 p < .01	+0,5 -2,0 - +3,0 p < .01	+0,79 0 - +4,0 p < .05	+0,34 -3,0 - +3,0 p < .001
Test 13	+0,09 -5,0 - +5,0 n.s.	-0,35 -7,0 - +3,0 n.s.	-0,57 -5,0 - +2,0 n.s.	+0,30 -4,0 - +5,0 n.s.
Test 14	+0,13 -2,0 - +3,0 n.s.	0 -3,0 - +2,0 n.s.	-0,07 -2,0 - +2,0 n.s.	-0,42 -3,0 - +2,0 p < .01

13-	-0,26	-0,94	-1,57	-0,16
	-4,0 - +3,0	-7,0 - +4,0	-5,0 - +1,0	-6,0 - +8,0
	n.s.	p < .05	p < .02	n.s.
D-Wert	-0,72	+0,22	+1,39	+0,67
	-5,0 - +3,5	-3,0 - +3,25	-0,75 - +8,25	-6,5 - +6,25
	p < .01	n.s.	.02 > p > .01	p < .01
Rechtschreibung SW (Standardwert)	+1,75	+2,24	+4,07	+4,16
	-7,0 - +11,0	-16,0 - +16,0	-8 - +26	-8 - +20
	p < .01	n.s.	p < .10	p < .01

p < .01 = Die Unterschiede zwischen beiden Messungen
sind nur in 1% der Fälle zufällig

n.s. = Unterschiede nicht signifikant

.10 > p > .05 = nur in 5% bis 10% der Fälle sind die Unterschiede zufällig

Die Tabelle 2 kann so gelesen werden:

Für Stichprobe 1 war bei der Testwiederholung für den Gesamtwert ein Zugewinn von 1,08 c- Werten festzustellen. Diese Unterschiede waren auf dem 0.01% - Niveau signifikant, der Zugewinn war also nicht zufällig. Allerdings reichten die Veränderungen in dieser Gruppe von einem Zugewinn von + 3,8 bis – 1. D. h.: der höchste Zugewinn betrug bei einer Person 3,8 c – Werte gegenüber dem ersten Test, während eine andere Person sich gegenüber dem ersten Test beim zweiten Test sogar um einen c- Wert *verschlechterte!*

6 . Vergleich guter und schlechter Lehrgangsteilnehmer

Unterscheiden sich gute von schlechten Lehrgangsteilnehmern hinsichtlich der Veränderung ihrer Testwerte? Tabelle 3 zeigt dazu die Werte von Lehrgangsteilnehmern mit der Abschlussnote 2 und denjenigen, die den Lehrgang nicht bestanden.

Interessant ist – wie die ersten Zahlenreihen zeigen - auch folgendes Ergebnis: Bei der *Gruppe* der guten Lehrgangsteilnehmer nahm die Streuung ($S = - 3,49 \sigma$) der Gesamtpunktzahl ab, während sie bei den schlechten etwa gleich blieb ($S = + 0,11 \sigma$). Die Leistungen der guten Lehrgangsteilnehmer wurden also homogener, ihre Gesamtpunktzahlen lagen enger beieinander.

8. Die Stabilität des Auswahlverfahrens

Die Qualität des Auswahlverfahrens zeigt sich unter anderem auch darin, dass die erzielten Werte relativ stabil sind. Dazu gibt es bezüglich des hier angesprochenen Auswahlverfahrens ein „natürliches Experiment“: Einige abgelehnte Bewerber mussten bei erneuter Bewerbung das Auswahlverfahren wiederholen. Die Ergebnisse für diese Gruppe von 18 Personen sind in den Tabellen 4, 5 und 6 dargestellt. Es zeigt sich, dass dieses „natürliche Experiment“ die Ergebnisse der vorher erwähnten Untersuchungen bestätigt.

Tabelle 4 zeigt z.B., dass bei den Untertests 1+2 nur ein durchschnittlicher Zuwachs von 0,61 c- Werten erzielt wurde, bei 10 Bewerbern veränderten sich die Werte für 1+2 beim 2. Auswahlverfahren nicht, bei 6 wurden ein Zugewinn von 1 c- Wert, bei einem Bewerber von 2 c- Werten und bei einem Bewerber von +3 c- Werten erzielt.

Tabelle 5 zeigt, wie individuell sich die Werte bei einigen Bewerbern veränderten.

Tabelle 6 zeigt die Veränderung der D- Werte für die Bewerber.

Literatur:

Füllgrabe, U. (2002). *Psychologie der Eigensicherung. Überleben ist kein Zufall.*

Stuttgart: R. Boorberg Verlag

Greif, S. (1972). *Gruppenintelligenztests.* Bern: Herbert Lang