Zur Interpretation objektiver Testresultate

Zur Interpretation einfacher Tests

Jeder statistische Kennwert ist mit einer gewissen Unsicherheit behaftet → Standard- oder Stichprobenfehler

Der *Standardfehler* einer Statistik ermöglicht die Beantwortung folgender Fragen:

1. Ob eine bestimmte Statistik signifikant ist, d.h. ob sie mit Sicherheit vom Wert Null abweicht.
   → mit t-Test
2. Innerhalb welcher Grenzen der Parameter, also die der Population entsprechende Statistik, zu erwarten ist.
   → Vertrauensgrenzen CL (confidential limits), der Bereich, innerhalb dessen bei einer vorgegebenen Irrtumswahrscheinlichkeit (5%/1%) der Parameter, d.h. der „wahre Kennwert“ erwartet werden darf.
   → Dass man den wahren Testwert nicht genau bestimmen kann, liegt an der mangelnden Reliabilität des Tests

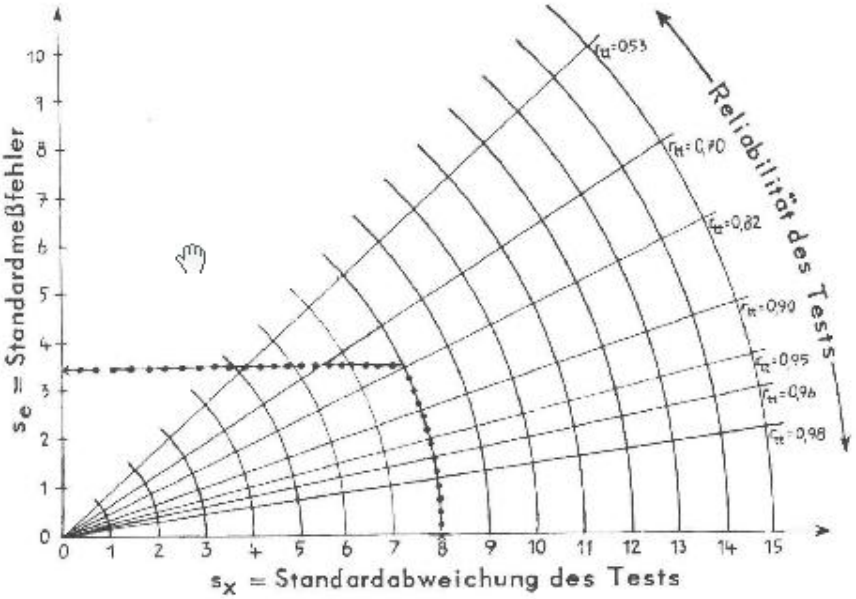


Abb. 15.1: Nomogramm der Beziehungen zwischen Reliabilität r_{tt}, Standardabweichung s_X und Standardmeßfehler s_e.

Der **Standardmessfehler** ist derjenige Anteil an der Standardabweichung eines Tests, der zu Lasten seiner Unreliabilität geht.

Rechenbeispiel Nr. 15.1:
Nehmen wir an, ein Intelligenztest habe eine Standardabweichung von $s_X = 7{,}0$ und eine Reliabilität von $r_{tt} = 0{,}91$. Was kann man über die Vertrauenswürdigkeit eines beliebigen Rohwertes dieses Tests aussagen? Nach Formel (15.4) ist

$$s_e = 7{,}0\sqrt{1 - 0{,}91} = 2{,}1\,,$$

so daß der wahre Rohwert eines Pbi nach Formel (15.5)

$$CL_X = X_i \pm 1{,}96 \cdot 2{,}1 = X_i \pm \sim 4$$

mit 5%iger Irrtumswahrscheinlichkeit höchstens 4 Rohwertpunkte über oder 4 Rohwertpunkte unter dem jeweils gemessenen Rohwert X_i liegen wird.

- Die Standardmessfehler verschiedener Tests sind untereinander vergleichbar, wenn man ihr Streuungsmaß vereinheitlicht (Normierung)
- Die Bestimmung eines Vertrauensintervalls mittels des Standardmessfehlers ist nur dann ganz richtig, wenn man vom wahren Wert T_i auf den beobachteten Testwert X_i schließen würde, nicht jedoch umgekehrt. Grund: beobachteter Wert und Messfehler sind nicht unkorreliert in KTT.

- Nach *Schmolck* ist es besser, den wahren Wert und das zugehörige Vertrauensintervall über *regressionsanalytischen Ansatz* zu bestimmen.
  - Der geschätzte wahre Wert T_i ist nur im Falle von $r_{tt} = 1$ gleich dem Testwert X_i, sonst weicht er um einen gewissen, durch die Reliabilität bestimmten Betrag ab
  - Hier ergibt sich ein etwas kleineres Vertrauensintervall für den wahren Wert, als bei dem herkömmlichen Ansatz

Aber: für diagnostische Praxis reicht der herkömmliche Ansatz aus; der regressionsanalytische Ansatz sollte höchstens in Einzelfällen bei sehr extremen Testergebnissen herangezogen werden.

Die Beurteilung interindividueller Unterschiede

Unterscheiden sich zwei Pbn mit wenig unterschiedlichen Resultaten in dem gleichen Test *tatsächlich* ihrer Leistung nach?

Prüfmöglichkeiten:

- Konfidenzintervalle beider Testwerte
  - Keine Überschneidung = beide Pbn sind unterschiedlich leistungsfähig
  - Überschneidung = Beibehaltung der Null-Hypothese (Annahme, dass kein Unterschied zwischen den Pbn besteht)
- Z-Test
  - Sehr scharf, aber auch aufwendig
- Kritische Differenz
  - Überschreitung drückt statistische Bedeutsamkeit aus
  - Gilt nur für eine definierte Irrtumswahrscheinlichkeit (meist 5% mit zugehörigem z-Wert von 1,96; 1%≙2,58)

Rechenbeispiel Nr. 15.3:
Es sei gefragt, wie groß die kritische Differenz – ausgedrückt in Einheiten des IQ – bei einem allgemeinen Intelligenztest mit $r_{tt} = 0{,}93$ sei, wenn wir einen sehr strengen Maßstab mit $P = 1\%$ anlegen wollen.

$$(IQ_1 - IQ_2)_{0,01} = 2{,}58 \cdot 15 \sqrt{2(1 - 0{,}93)} = 14{,}5$$

Wir stellen also fest: Zwei Pbn, die sich in dem besagten Test um fast 15 IQ-Einheiten unterscheiden, sind mit nahezu absoluter Sicherheit verschieden intelligent. Dabei setzen wir naturgemäß die Validität als logisch begründet voraus; nur unter diesem Aspekt ist die Aussage in ihrer Formulierung zutreffend. Wollten wir uns mit der geringeren Sicherheit von $P = 5\%$ begnügen, so würden bereits Unterschiede von 11 und mehr IQ-Punkten bedeutsam sein.

Die Beurteilung intraindividueller Unterschiede

- Pb wird in einer Zeitspanne mit demselben Test oder seiner Parallelform erneut untersucht
- Unter der Voraussetzung, dass kein Wiederholungsgewinn zu erwarten ist, lässt sich feststellen, ob sich das fragliche Persönlichkeitsmerkmal verändert hat
- Bei Interesse an Wiederholungsgewinn nehmen wir an, dass das Merkmal *konstant* ist
- Auch hier ist z-Test möglich

Sinn der Nullhypothese: Unterschiede, die nicht gesichert sind, können trotzdem bestehen, nur ist die Wahrscheinlichkeit nicht hoch genug, als dass man auf sie vertrauen könnte.

Rechenbeispiel Nr. 15.4:
Ein Pb wurde mit einem Intelligenztest getestet. Nach 14 Tagen wurde der Test wiederholt. Die beiden Ergebnisse waren $IQ_1 = 110$ und $IQ_2 = 115$. Darf man die Differenz im Sinne eines Wiederholungsgewinns interpretieren?
Wenn wir eine Reliabilität von $r_{tt} = 0{,}92$ annehmen und $P = 5\%$ festlegen, dann ergibt sich die kritische Differenz

$$(IQ_2 - IQ_1)_{0.05} = 1{,}96 \cdot 15 \cdot \sqrt{2(1 - 0{,}92)} = 12 .$$

Man darf den Leistungszuwachs also nicht interpretieren, da er kleiner als die kritische Differenz ist.

Verlaufsprofil = ein Pb wird in regelmäßigen Zeitabständen mittels einer Testserie auf den veränderlichen Ausprägungsgrad eines Persönlichkeitsmerkmals hin überprüft.
Statt Testserie kann man auch verschiedene Tests mit gleichem Validitätsanspruch verwenden.

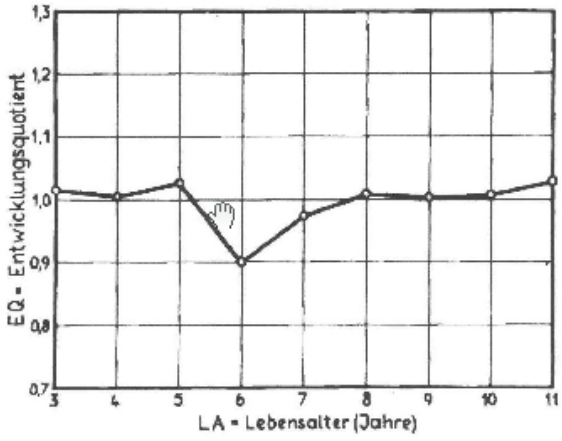


Abb. 15.2: Beispiel eines Verlaufsprofils des Entwicklungsquotienten (EQ).

Es können auch allmähliche Veränderungen, z.B. ein linearer Anstieg des EQ, teststatistisch beurteilt werden.

Ist Abweichung bedeutsam?

Vom letzten gesetzmäßigen Testwert ausgehend die z-Werte kumulativ summieren und beim Überschreiten von 1,96 bzw. 2,58 die reale Veränderung definitiv feststellen

Rechenbeispiel Nr. 15.5:

Halten wir uns an Abb. 15.2 und nehmen wir an, ein Kind hätte vom 3. bis zum 5. Lebensjahr ziemlich konstant einen $\overline{EQ} = 1{,}01$ gezeigt. Wir fragen, ob der Einschnitt bei 6 Jahren (EQ = 0,90) bedeutsam ist; σ beträgt für den EQ 0,1, die Reliabilität des Entwicklungstests wollen wir mit $r_{tt} = 0{,}75$ und P = 5% annehmen. Wir berechnen zunächst den z-Wert für die Differenz zwischen der 6-Jahresleistung und dem Mittel der vorhergehenden Leistungen nach Formel (15.13)

$$z_1 = \frac{1{,}01 - 0{,}90}{0{,}1\sqrt{2(1 - 0{,}75)}} = 1{,}56\,.$$

Der beobachtete Leistungsabfall ist statistisch nicht bedeutsam. Beziehen wir die 7-Jahresleistung noch mit ein, so bilden wir den z-Wert der Differenz von EQ = 1,01 zu EQ = 0,975.

$$z_2 = \frac{1{,}01 - 0{,}975}{0{,}1\sqrt{2(1 - 0{,}75)}} = 0{,}50$$

Zur Beurteilung des gesamten Einschnittes bei 6 Jahren addieren wir z_1 und z_2 algebraisch: 1,5 + 0,5 = 2,0. Der Entwicklungseinschnitt ist also bedeutsam und entspricht offenbar einer vorübergehenden Retardation.

Beurteilung weiterer Unterschiede

Überprüfung des Unterschieds in den Mittelwerten eines Tests, der in zwei Gruppen von Pbn durchgeführt worden ist, unter Berücksichtigung der Reliabilität auf Signifikanz (*Griesang*)

Rechenbeispiel Nr. 15.6:

Zwei Stichproben von jeweils 10 Jungen und 10 Mädchen im Alter von 12 Jahren wurden mit einem in T-Werten normierten Wortschatztest untersucht. Die Jungen hatten einem Mittelwert von $\bar{X} = 52$, die Mädchen einen solchen von $\bar{X} = 55$. Darf man diesen Unterschied unter Berücksichtigung der Reliabilität des Tests von nur $r_{tt} = 0{,}8$ bei einer Irrtumswahrscheinlichkeit von 5% interpretieren?

Es ergibt sich die folgende kritische Differenz:

$$(\bar{X}_1 - \bar{X}_2)_{0,05} = 1{,}96 \cdot 10 \cdot \sqrt{\frac{2}{10}(1 - 0{,}8)} = 4\,.$$

Die tatsächliche Differenz von 3 ist kleiner als die kritische Differenz, der Unterschied zwischen den beiden Gruppen darf also nicht interpretiert werden.

Durch kleine Veränderungen der Grundformel können auch folgende Dinge verglichen werden

- Mittelwert einer Gruppe mit einem vorgegebenen Normwert
- Testergebnisse eines einzelnen Pbn mit dem Mittelwert einer Gruppe
- Vergleich der Testergebnisse von zwei Pbn
- Ob das Testergebnis eines Pbn von einem bestimmten Normwert abweicht

Rechenbeispiel Nr. 15.7:
Ein Abiturient hat in einem Intelligenztest einen IQ von 122 erreicht. Der Mittelwert „aller" Abiturienten liegt bei 112. Darf man diesen Unterschied interpretieren ($r_{tt} = 0{,}91$, $P = 5\%$)?
Nach Formel (15.21) ist die kritische Differenz

$$(X_i - M)_{0,05} = 1{,}96 \cdot 15 \sqrt{1 - 0{,}91} = 9 .$$

Die beobachtete Differenz ist mit 10 IQ-Punkten größer als die kritische Differenz, deshalb darf der Unterschied interpretiert werden.

Zur Interpretation von Testprofilen

Was ist ein Testprofil?

- Resultiert, wenn individuelle Testwerte aus einem Test oder Inventar, in dem mehrere Konstrukte erfasst werden, gemeinsam betrachtet und z.B. in ein Profildiagramm übertragen werden.
- Lassen sich z.B. in Tests, die verschiedene Intelligenzfacetten erfassen oder in mehrdimensionalen Persönlichkeitsinventaren (z.B. zur Erfassung der Big Five) erstellen.
- Interesse daran, ob ein individuelles Testprofil über die Zeit stabil geblieben ist oder sich verändert hat

Dazu müssen zwei Faktoren berücksichtigt werden:

- Die Profilgestalt
- Die Profilhöhe

Profilgestalt

- Betrifft die relativen Positionen der Profilmerkmale zueinander, also deren Rangreihe
- Ipsativ-differentielle Stabilität bzw. Veränderung = Stabilität/Veränderung der relativen Positionen der Profilmerkmale über die Zeit
- Hierfür würde es reichen, eine Q-Korrelation über die individuellen oder gruppenbezogenen Profilmerkmale zu berechnen. Aber dabei wird Profilhöhe nicht berücksichtigt

Profilhöhe

- Bezieht sich auf die möglichen absoluten Differenzen zwischen den einzelnen wiederholt gemessenen Profilmerkmalen = ipsativ-absolute Stabilität/Veränderung

Die Bedeutung des Profilreliabilitätskoeffizienten

Profilreliabilität

- impliziert die Interkorrelation der Einzeltests
- Am größten, wenn die Einzeltests hoch reliabel sind und zugleich niedrig interkorrelieren

Aber: ein Profil, dessen Interkorrelationen ebenso hoch sind wie dessen Einzelreliabilitäten kann auch nur ein *Scheinprofil* sein → Testserie, deren Einzeltests allesamt dasselbe Persönlichkeitsmerkmal untersuchen.

Rechenbeispiel Nr. 15.8:
Ein Viererprofil mit den Reliabilitätskoeffizienten $r_{11} = 0{,}87$; $r_{22} = 0{,}91$; $r_{33} = 0{,}94$; $r_{44} = 0{,}88$ und den Interkorrelationen $r_{12} = 0{,}34$; $r_{13} = 0{,}40$; $r_{14} = 0{,}25$; $r_{23} = 0{,}31$; $r_{24} = 0{,}20$ und $r_{34} = 0{,}60$ hätte nach Formel (14.2) eine Profilreliabilität von

$$_{prof}r_{tt} = \frac{0{,}90 - 0{,}35}{1 - 0{,}35} = 0{,}85 .$$

Da wir an die Profilreliabilität nicht so hohe Anforderungen stellen wie an die Reliabilität standardisierter Einzeltests, können wir mit dem differentialdiagnostischen Wert dieses Profils sehr zufrieden sein.

Reliabilitätskoeffizient

- vermittelt nur einen allgemein orientierten Eindruck.
- Ist unbrauchbar zur Interpretation von Profildifferenzen.
- Diagnositische Wert eines Profils wird eher verschleiert als erklärt, da alle Werte in einen Topf geworfen werden
- Über .8 = hoch; zwischen .6 und .8 = befriedigend

Die speziellen Aussagemöglichkeiten eines Testprofils

- Ein Profil kann ebenso wie mehrere voneinander unabhängig durchgeführte Einzeltests interpretiert werden
- Die Eigenart eines Profils liegt aber in der *Vergleichbarkeit* der Einzeltests
  - ✧ Durch die einheitliche Standardskala gewährleistet → differentielle Interpretation der Profilunterschiede
  - ✧ Kardinalfrage vor Interpretation: Ist das beobachtete Profil ein echtes Profil oder ein Scheinprofil, d.h. sind die beobachteten Unterschiede so ausgeprägt, dass sie nicht durch Zufall entstanden sein können?

Statistische Profilinterpretation:

- Über Konfidenzintervalle der Einzeltests
- Für die Deutbarkeit einer Profildifferenz wird gefordert, dass sich die Konfidenzintervalle der beteiligten Tests nicht überschneiden
- Fasst man die einander zugewandten Hälften der beiden Konfidenzintervalle zusammen, erhält man den *kritischen Unterschied* für je zwei Testpunktwerte eines Profils eines Pb

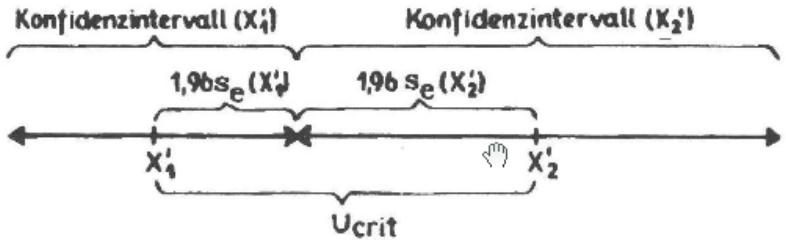


Abb. 15.3: Die Bedeutsamkeit des Unterschiedes zweier Testpunktwerte, beurteilt nach dem Konfidenzintervall.

Rechenbeispiel Nr. 15.9:

Wollten wir wissen, wie groß die kritische Differenz für den ZR-Test (induktives Zahlendenken) und den WÜ-Test (räumliches Vorstellenkönnen) im IST-AMTHAUER (1955) ist, so müßten wir zunächst die beiden Reliabilitätskoeffizienten nachlesen: $r_{tt}(ZR) = 0{,}96$ und $r_{tt}(WÜ) = 0{,}89$. Die Standardwerte des IST haben eine Streuung von $s = 10$. Begnügen wir uns mit einer Irrtumswahrscheinlichkeit von 5%, so ist $z_{crit} = 1{,}96$. Die kritische Differenz berechnet sich also nach Formel (15.23) zu:

$$(ZR_i - WÜ_i)_{0,05} = 1{,}96 \cdot 10 \left(\sqrt{1 - 0{,}96} + \sqrt{1 - 0{,}89}\right) = 10{,}4 .$$

Die Profilpunkte in ZR und WÜ müßten also mehr als 10 SW-Punkte auseinanderliegen, ehe wir behaupten dürften, daß induktives Zahlendenken bei einem bestimmten Pb_i besser ausgebildet sei als räumliches Vorstellenkönnen oder umgekehrt.

Die Interpretation von Profildifferenzen kann wesentlich präziser und ökonomischer gestaltet werden, wenn man dem Konzept des *Standardmessfehlers von Profildifferenzen* folgt (LIENERT; 1957)

$$s_{ed} = s_X \sqrt{2 - (r_{11} + r_{22})} . \tag{15.24}$$

Der auf s_{ed} gegründete kritische Unterschied – mit d_{crit} symbolisiert – errechnet sich nach der Beziehung

$$d_{crit} = z_{crit} \cdot s_{ed} = z_{crit}\, s_x \sqrt{2 - (r_{11} + r_{22})} . \tag{15.25}$$

Zweckmäßigerweise stellt man sich die kritischen Differenzen in einer *Matrix* übersichtlich dar, wie dies in Beispiel Nr. 15.10 geschehen ist.

Rechenbeispiel Nr. 15.10:

Das Testmanual des IST-AMTHAUER (1955) nennt folgende Halbierungskoeffizienten für die Einzeltests: $r_{tt}(SE) = 0{,}91$; $r_{tt}(WA) = 0{,}88$; $r_{tt}(AN) = 0{,}86$; $r_{tt}(GE) = 0{,}93$; $r_{tt}(ME) = 0{,}90$; $r_{tt}(RA) = 0{,}92$; $r_{tt}(ZR) = 0{,}96$; $r_{tt}(FA) = 0{,}84$ und $r_{tt}(WÜ) = 0{,}89$. Die Reliabilität des Gesamt- bzw. Durchschnittpunktwertes beträgt $r_{tt}(TOT) = 0{,}967$; die Streuung der Profilnormen des IST ist $\sigma = 10$.

Rechnen wir uns nun die kritischen Differenzen auf der 5%-Stufe für jedes Testpaar nach Formel (15.25) aus; wir bringen als Beispiel nur die Berechnung der Differenz SE – WA:

$$d_{crit}(SE - WA) = 1{,}96 \cdot 10 \sqrt{2 - (0{,}91 + 0{,}88)} = 9{,}0 .$$

Die Werte tragen wir in die rechte obere Hälfte einer Profildifferenzenmatrix ein.

Tabelle 15.2

Test	SE	WA	AN	GE	ME	RA	ZR	FA	WÜ	TOT
SE	8,3	9,0	9,4	7,8	8,5	8,1	7,1	9,8	8,8	6,8
WA		9,6	10,0	8,5	9,2	8,8	7,8	10,4	9,4	7,6
AN			10,4	9,0	9,6	9,2	8,3	10,7	9,8	8,1
GE				7,3	8,1	7,6	6,5	9,4	8,3	6,5
ME					8,8	8,3	7,3	10,0	9,0	7,1
RA						7,8	6,8	9,6	8,5	6,5
ZR							5,5	8,8	7,6	5,8
FA								11,1	10,2	8,5
WÜ									9,2	7,3
TOT										3,3

Benötigen wir auch die kritischen Differenzen für die Beurteilung von Wiederholungsergebnissen oder für den Vergleich zweier Pbn in dem gleichen Einzeltest, so rechnen wir -- wiederum für P = 5% -- nach Formel (15.14):

$$d'_{crit}(SE) = 1{,}96 \cdot 10 \sqrt{2(1 - 0{,}91)} = 8{,}3 .$$

Die erhaltenen Werte haben wir in die Diagonalfelder der Tab. 15.2 eingetragen.

Nach der Differenzenmatrix ergeben sich beispielsweise folgende Interpretationsmöglichkeiten:

1. Zwei Pbn besitzen nur dann mit Sicherheit eine unterschiedliche Merkfähigkeit (ME), wenn der Profilwert des einen mindestens um 9 (8,8) SW-Punkte höher liegt als der des anderen.
2. Ein Pb zeigt bei einem Wiederholungsversuch mit der Parallelform des IST unter anderen Bedingungen dann eine signifikant bessere Leistung im praktischen Rechnen, wenn das zweite Resultat im RA-Test mindestens 8 (7,8) SW-Punkte besser ist als das erste Resultat. Wobei wir allerdings voraussetzen, daß der zwischen den Versuchen liegende Zeitraum groß genug ist, um einen Übungsgewinn zu vermeiden.
3. Ein Pb hat dann ein besseres Raumvorstellungs- als induktives Denkvermögen, wenn sein WÜ-Wert mindestens 8 (7,6) SW-Punkte über seinem ZR-Wert liegt. Nach dem weniger präzisen Konfidenzintervallkonzept war hierzu eine Differenz von $U_{crit} = 10{,}4$ SW-Punkten erforderlich.

Die Beurteilung von Profildifferenzen wird wesentlich einfacher und ökonomischer, wenn man anstelle der Differenzenmatrix eine „globale“ kritische Differenz im Testmanual als Interpretationshilfe angibt.

Eine allgemeine Formel zur Prüfung von Profilunterschieden

Der Formelapparat von Kristof (1958)

- Berücksichtigt Profilgestalt und Profilhöhe bei der Stabilitätsbestimmung
- Zwei individuelle oder gruppenbezogene Profile werden zuerst auf „Deckungsgleichheit“ getestet (globaler Profilvergleich)
    - Wenn sign. globaler Profilunterschied weitere Prüfung:
        - ♦ Basiert Diskrepanz auf Unterschieden in Profilhöhe und/oder Profilgestalt?

Um das zu überprüfen, müssen zwei Voraussetzungen erfüllt sein:

1) Die Ausprägungen der Profilmerkmale müssen vor der Verrechnung auf der Basis eines Eichstichprobe normiert werden.
2) Die (differentiellen) Stabilitäten (Retest-Korrelationen) der Profilmerkmale müssen bekannt sein.

KRISTOF (1958) hat einen allgemeinen χ^2-Test zur Prüfung des globalen Unterschieds zwischen zwei *Gruppenprofilen* angegeben.

$$\chi^2 = \frac{N_1 N_2}{(N_1 + N_2)\, s_X^2} \sum_i \frac{D_i^2}{1 - r_{ii}} \qquad df = k$$

Aus der Grundformel lassen sich verschiedene Spezialfälle ableiten.

z.B. Vergleich
- eines durchschnittlichen Profils einer Gruppe mit einem Normprofil
-eines individuellen Testprofils eines Pb mit dem Durchschnittsprofil einer Gruppe
-eines individuellen Profils mit einem Normprofil
-der individuellen Testprofile von zwei Pbn

Rechenbeispiel Nr. 15.11:
Ein Gymnasiast möchte Jura studieren. Bei einer Studienberatung erzielte er im IST-AMTHAUER die in Tabelle 15.3 zusammengestellten Werte. Es ist zu prüfen, ob sich sein Testprofil von dem Normprofil von Juristen bei einem vorgegebenen Risiko von 5% unterscheidet. Tabelle 15.3 enthält auch dieses Normprofil, sowie die aus Tabelle 15.2 entnommenen Reliabilitätskoeffizienten und die einzelnen Zwischenergebnisse, die bei Verwendung von Formel (15.31) anfallen. Dabei ist $Q = D_j^2/(1 - r_{tt})$. Q wird nur auf ganze Zahlen genau bestimmt.

Tabelle 15.3

Test	SE	WA	AN	GE	ME	RA	ZR	FA	WÜ	
RW	105	110	113	115	100	110	112	110	105	
Norm	110	108	111	115	106	103	107	103	98	
r_{tt}	0,91	0,88	0,86	0,93	0,90	0,92	0,96	0,84	0,89	
D_t^2	25	4	4	100	36	9	25	49	49	
$1 - r_{tt}$	0,09	0,12	0,14	0,07	0,10	0,08	0,04	0,16	0,11	$\sum$
Q	278	32	29	00	360	113	625	306	445	2188

$$\chi^2 = \frac{1}{100} \cdot 2188 = 21{,}88\,.$$

In Tafel 5 des Anhangs findet man bei 9 Freiheitsgraden und P = 5% einen kritischen Wert von 16,9. Die Nullhypothese muß verworfen werden, die beiden Profile unterscheiden sich.

Die Ähnlichkeitsbeurteilung von Profilen

Durch Korrelation zweier Profile → möglich, aber man lässt die Tatsache außer Acht, dass die Höhe des Profils praktisch eine Rolle spielt

Ähnlichkeitsindex D (OSGOOD und SUCI, 1952) → berücksichtigt Profilhöhe und die Verlaufsgestalt bei der Ähnlichkeitsbeurteilung gleichermaßen.

$$D = \sqrt{\sum d_t^2}$$

In dieser Formel bedeutet:

d_t = Profildifferenz zweier Pbn in einem beliebigen Test t.

Rechenbeispiel Nr. 15.12:
Angenommen, es hätten zwei Pbn folgende Fragebogenprofile (in T-Werten) erzielt, und diese sollten mit einem Berufsprofil – etwa dem für soziale Berufe – verglichen werden. Die Daten enthält Tab. 15.4.

Tabelle 15.4

	Extra-version	Domi-nanz	Neuroti-zismus	Soziale Einstellung
Pb A	43	39	64	70
Pb B	56	52	45	67
Berufs-profil	51	40	57	65

Wir möchten wissen, welches der beiden Individualprofile A und B dem Berufsprofil P näher steht. Zu diesem Zweck berechnen wir die beiden Ähnlichkeitsindizes D_{AP} und D_{BP} nach Formel (15.32).

$$D_{AP} = \sqrt{(43-51)^2 + (39-40)^2 + (64-57)^2 + (70-65)^2} = 11{,}8$$

$$D_{BP} = \sqrt{(56-51)^2 + (52-40)^2 + (45-57)^2 + (67-65)^2} = 17{,}8\,.$$

Das Profil des Pb A ähnelt also dem typischen Berufsprofil für soziale Berufe mehr als das Profil des Pb B. In gleicher Weise könnten wir auch noch den Ähnlichkeitsindex der beiden Individualprofile zueinander berechnen.

Ähnlichkeitsindex D

- Ist nur ein Vergleichsmaß
- Hängt von der Wahl der Profilskale und den Interkorrelationen ab
- Möglichkeit ihn zu standardisieren: in einen Korrelationskoeffizienten umrechnen

Rechenbeispiel Nr. 15.13:
Würden wir die beiden Ähnlichkeitsindizes D_{AP} und D_{BP} aus Beispiel Nr. 15.12 in Korrelationskoeffizienten umrechnen, so würden sich nach Formel (15.33) folgende Werte ergeben:

$$r_{PA} = \frac{2 \cdot 3{,}36 \cdot 10^2 - 11{,}8^2}{2 \cdot 3{,}36 \cdot 10^2 + 11{,}8^2} = 0{,}66$$

$$r_{PB} = \frac{2 \cdot 3{,}36 \cdot 10^2 - 17{,}8^2}{2 \cdot 3{,}36 \cdot 10^2 + 17{,}8^2} = 0{,}36\,.$$

Wir erhalten also für den Ähnlichkeitsindex D_{AP} einen wesentlich höheren Korrelationskoeffizienten als für den Ähnlichkeitsindex D_{BP}.

Die Höhe des Korrelationskoeffizienten wird wie üblich interpretiert. Korrelationen $>0{,}8$ gelten als sehr groß, solche $>0{,}6$ als groß und solche zwischen 0,6 und 0,4 als mittelhoch.

Wenn man für zwei beliebige Profile sowohl den Unterschied mittels eines χ^2-Tests überprüft als auch den Grad der Übereinstimmung mit dem Profilähnlichkeitskoeffizienten r_p nach CATTELL (Bsp. 15.13) bestimmt hat, kann man den Zusammenhang zwischen diesen beiden Kennwerten nach HUBER auf folgende einfache Weise darstellen:

$$r_p = \frac{\chi^2_{0,50} - \chi^2}{\chi^2_{0,50} + \chi^2} \qquad df = k$$

Diese Gleichung bietet die Möglichkeit, für einen beliebigen Profilvergleich r_p zu bestimmen, also z.B. für den Vergleich von

- Zwei Gruppenprofilen
- Zwei Gruppenprofilen bei gleichem N
- Einem Gruppen- mit einem Normprofil
- Einem individuellen mit einem Gruppenprofil
- Einem individuellen mit einem Normprofil
- Zwei individuellen Profilen

Rechenbeispiel Nr. 15.14:

In Rechenbeispiel 15.11 war das Profil eines Gymnasiasten mit dem aus $k = 9$ Subtests bestehenden Normprofil von Juristen verglichen worden. Es hatte sich $\chi^2 = 21{,}88$ ergeben. Nach Tafel 5 im Anhang ist $\chi^2_{0,50} = 8{,}34$. Nach Formel (15.34) errechnet sich daraus ein Profilähnlichkeitskoeffizient r_p von

$$r_p = \frac{8{,}34 - 21{,}88}{8{,}34 + 21{,}88} = -0{,}45 .$$

Die beiden Profile stehen in einem mittleren, aber gegenläufigen Zusammenhang, was man bei einer Inspektion der Werte in Tab. 15.3 leicht erkennen kann.

Die Kriteriumsvorhersage bei Tests und Testbatterien

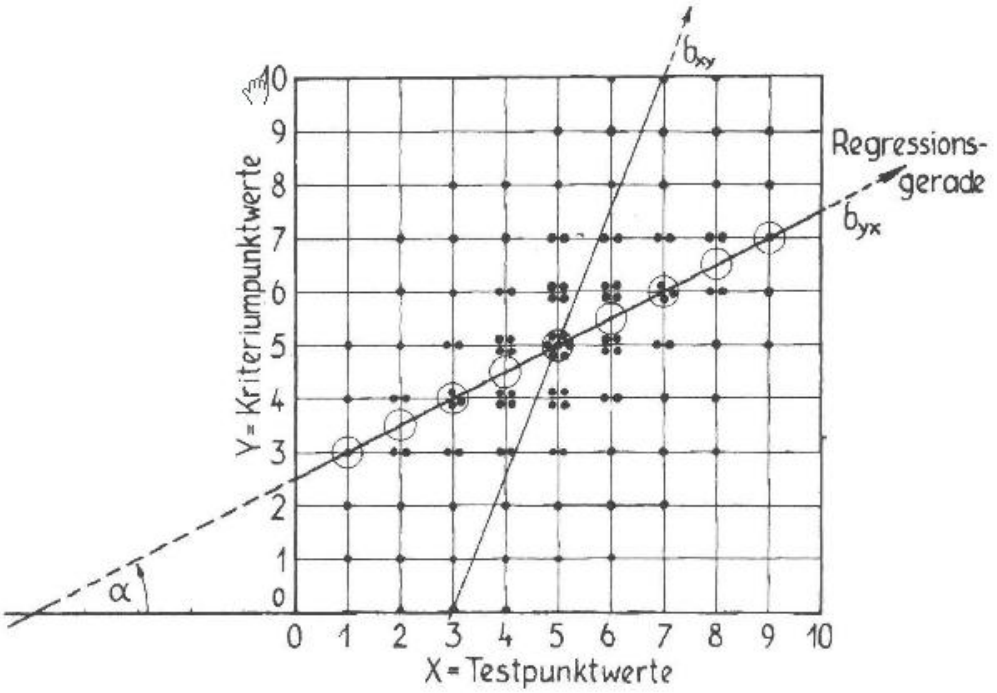
Die einfache Regression bei der Interpretation empirisch valider Tests

Stehen zwei Messwertreihen wie Test- und Kriteriumswerte in linearer Abhängigkeit, so kann man aufgrund der Regressionsgleichung zu einem beliebigen Testpunktwert den entsprechenden Kriteriumswert voraussagen.

Test wurde an 95 Pbn extern validiert.
Wie wird sich ein Pb mit 7 Testpunkten hinsichtlich derjenigen Tätigkeit, die der Test messen soll, bewähren, wenn für den Grad der Bewährung bis zu 10 Punkten vergeben werden?

→ *Kriteriumserfolg der 13Pbn, die 7 Testpunkte bei der Validierung erreicht haben, ist sehr unterschiedlich → Entscheidung für Mittelwert (3 von 13 = 23%; mit Irrtumswahrscheinlichkeit sind es schon 54%)*

(umgekehrt: zu jedem Kriteriumwert einen Testwert vorhersagen = Regression der X-Werte auf die Y-Werte)



(Verbindung aller Mittelwerte = Regression der Y-Werte auf die X-Werte)

Abb. 15.4: Regression der Kriteriumpunktwerte in Bezug auf die Testpunktwerte.

Woher stammt der Begriff „Regression" und was hat er zu bedeuten?
Vergleiche: (7/6) (9/7) (5/5) (1/3) → Pbn mit niedrigem Testwert haben gar nicht so ungünstige Erfolgsprognose → GALTON „Regression" → Rückschritt zur Mitte.

Rückschritt zur Mitte beruht darauf, dass viele unbekannte Variablen, wie z.B. Erfahrung, Interessen, Lernfähigkeit u.a., sowohl das Testergebnis als auch den Kriteriumwert beeinflussen. Somit tendieren diese eher zur Mitte.

Allgemein lässt sich die *lineare Regressionsgleichung* Y auf X folgendermaßen definieren:

vorhergesagte Kriterienwert eines Pb — *Testwert*

$$\tilde{Y}_i = b_{XY} \cdot X_i + K$$

Regressionskoeffizient (Anstieg der Regressionsgeraden → Tangens des Winkels α zwischen Regressionsgerade und X-Achse) gibt auch an, um welchen Betrag sich Y ändert, wenn X um eine Einheit weiterrückt (hier 0,5)

Achsenabschnitt (die Strecke, die die Regressionsgerade auf der Y-Achse abschneidet)

Wie kann man den Regressionskoeffizienten ohne Verwendung einer konkreten Korrelationstabelle berechnen?

Empirisch gegeben ist stets nur die Korrelation zwischen Test und Kriterium, der Validitätskoeffizient r_{tc}.
Unter Standardbedingungen (gleiche Streuungsmaße) könnten wir die Validität unseres Tests allein aus der Korrelationstabelle graphisch abschätzen: Tangens des Winkels α, also hier 0,5 als Validitätskoeffizient.

Bestimmungsformel für b_{yx}, wenn keine Standardbedingungen: $b_{yx} = r_{xy} \frac{s_y}{s_x}$

Wir standardisieren den Validitätskoeffizienten, indem wir ihn mit dem Faktor s_y/s_x multiplizieren.

Nach der folgenden Gleichung lässt sich für jeden beobachteten Testwert X_i ein Schätzwert Y_i voraussagen:

$$\tilde{Y}_i = r_{xy} \frac{s_y}{s_x} (X_i - \bar{X}) + \bar{Y}$$

Wenn Test- und Kriteriumwerte auf dieselbe Normenskala bezogen sind:

$$Z_{ci} = r_{tc}(Z_{ti} - \bar{Z}_t) + \bar{Z}_c$$

Es lohnt sich, eine Ablesetabelle herzustellen.
Die Beachtung des Regressionsprinzips wird – besonders bei wenig validen Testen – den Untersucher vor falschen Schlussfolgerungen bewahren.

Die multiple Regression bei der Kriteriumvorhersage

Regressionsprinzip wird auf die multiple Messung durch eine Testbatterie übertragen.
Statt der einfachen Regressionskoeffizienten werden *Partialregressionskoeffizienten* erster, zweiter usw. Ordnung verwendet.
Partialregressionskoeffizienten geben die mögliche Veränderung an, die der Einzeltest einer Testbatterie beim Kriterium bewirken würde, wenn die übrigen Tests keinen Einfluss hätten.
Der Einfluss aller Tests einer Batterie addiert sich algebraisch nach der multiplen Regressionsgleichung.

Allerdings verhilft man sich für Interpretationszwecke durch Ablesetabellen (S. 387)
Im Prinzip ist die multiple Korrelation von der einfachen Korrelation nicht verschieden. (hier Zusammenhang zwischen dem tatsächlich ermittelten Validitätskriterium und dem aufgrund mehrerer Tests vorausgesagten Validitätskriterium). Vorgehensweise:

- Testrohwerte in z-Werte transformieren
- Z-Werte mit den zugehörigen Betas gewichtet
- Die gewichteten z-Werte summiert

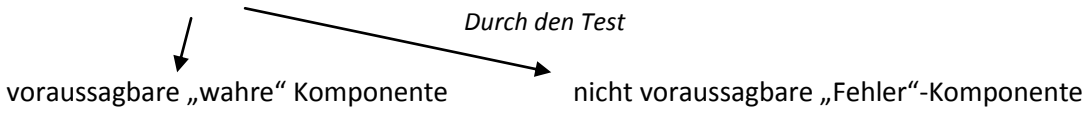
➔ Dann hat man in dieser Summe ein Äquivalent des vorausgesagten Validitätskriteriums

Wenn man dieses mit dem de facto erhobenen Validitätskriterium nach PEARSON korreliert, erhält man einen einfachen Validitätskoeffizienten, der mit dem multiplen über die Beta-Gewichte errechneten Validitätskoeffizienten identisch ist.

Die Unsicherheit der Kriteriumsvorhersage

Der Standardschätzfehler als Unsicherheitsmaß der Kriteriumsvorhersage

Varianz des Kriteriums:



Diese Fehlervarianz = s^2_{yx} ;
Wurzel daraus: s_{yx} = Standardschätzfehler *(Streuung der Punkte um die Regressionsgerade herum)*

Mittels des Standardschätzfehlers lässt sich die Unsicherheit der Kriteriumschätzung numerisch ausdrücken. Wir möchten wissen, wie genau eine Voraussage unter gegebenen Umständen ist, d.h. die Grenzen von Y angeben, innerhalb welcher mit einer vorgegebenen Sicherheit der „wahre“, bei späterer Bewährungskontrolle erhältliche Kriteriumwert liegen muss. → Konfidenzintervall.

Nur wenn ein Test einen hohen Validitätskoeffizienten r_{tc} und damit einen niedrigen sog. *Alienationskoeffizienten* $\sqrt{1-r}$ aufweist, ist die Voraussage eines Kriteriumswertes aufgrund eines Testwertes einigermaßen sicher bzw. vertrauenswürdig.
Selbst bei einem für extrem hohen Validitätskoeffizienten von 0,80 betrüge der Alienationskoeffizient noch 0,60.

➔ Es bedarf unrealistisch hoher Validitätskoeffiezienten, um individuelle Kriteriumsvoraussagen mit Sicherheit machen zu wollen!

Die Unsicherheit der Voraussage eines dichotomischen Kriteriums

Macht es dann überhaupt Sinn, individuelle Voraussagen zu treffen bzw. eine Ausleseentscheidung zu gründen?
Ja, weil nicht allein die Validität ausschlaggebend ist, sondern die Wirksamkeit der Auslese hängt auch vom Angebot und Bedarf an qualifizierten Arbeitskräften oder Ausbildungsplatzbewerbern ab.

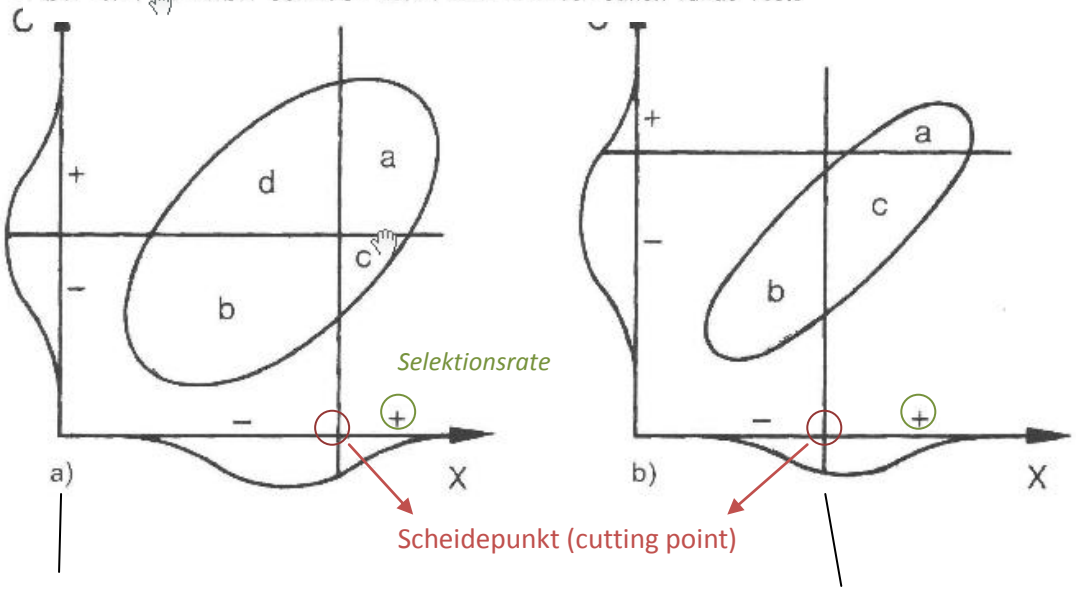
Also müssen wir folgende Faktoren beachten:

1. Die *Validität* des Tests im Hinblick auf eine unausgelesene Population potentieller Bewerber für einen Arbeitsplatz (in Betrieb, Schule und Ausbildungsstätte)
2. Die sog. *Selektionsrate*, d.h. den Anteil des Bedarfs im Verhältnis zum Angebot an Arbeitskräften aus der Bewerberpopulation
3. Den Anteil der *potentiell „Geeigneten"* in der unausgelesenen Bewerberpopulation (etwa zu ermitteln nach Einstellung einer Anzahl von Bewerbern ohne jegliche Testauslese).

Ein dichotomes Kriterium erhalten wir durch die Einteilung der Individuen in „geeignet" und „ungeeignet".

Zu1)

Abb. 15.5: Bewerber-Selektion durch zwei unterschiedlich valide Tests



Niedrige Validität hohe Validität

Senkrechte Linie trennt die ausgelesenen (+) von den zurückgewiesenen (-) Bewerbern
Waagrechte Linie trennt die potentiell Geeigneten (+) von den nicht geeigneten (-).

Man kann aus der Abbildung unmittelbar die Anteile der falschen und richtigen Entscheidungen ablesen:

- a und b = zu recht aufgenommenen bzw. abgelehnten Bewerber
- c = Anteil der Bewerber, die aufgenommen wurden, sich aber später nicht bewährt haben. Dieser Fehler geht zu Lasten der Institution.
- d = Bewerber, die abgelehnt werden, obwohl sie später erfolgreich gewesen wären. Dieser Fehler geht zu Lasten der Bewerber.

Bedeutung der Validität des Tests:

- Test mit $r_{tc} = 1$ erlaubt eine eindeutige Zuordnung, da sich Geeignete und Ungeeignete in ihren Testwertverteilungen nicht überlappen, wenn $r_{pbis} = 1$.
- Scheiderate sollte so gelegt werden, dass die Selektionsrate dem Anteil der Geeigneten entspricht; sonst Fehlentscheidungen trotz optimalem Test
- Ist $r_{tc} < 1$, liegt eine m.o.w. große Überlappung der Testwerte von Geeigneten und Ungeeigneten vor → Scheidepunkt so legen, dass Anteil der Fehlentscheidungen in beide Richtungen möglichst gleich sind (wenn von praktisch gleicher Konsequenz)

- Oft kann man aber den Scheidepunkt nicht frei wählen; somit verändern sich die Anteile der richtigen und falschen Entscheidungen
- Ist Validität gleich Null, ist keinerlei Voraussage möglich → Validität als notwendige, aber nicht allein hinreichende Voraussetzung für eine wirksame Eignungsvoraussage

Rechenbeispiel Nr. 15.15:

Mittels eines Schulreifetests wurden folgende Testwertverteilungen bei Eingeschulten (E) bzw. Schulreifen und Zurückgewiesenen (Z) bzw. Schulunreifen erhalten, wobei der Test keinen Einfluß auf die Entscheidung hatte (LIENERT und KROWARZ, 1959):

					SP										
E-Frequenz			3	3	3	8	14	25	35	61	52	48	38	22	12	5
Z-Frequenz	2	5	8	7	↓	8	4	7	9	3	1		1			
Testwertklasse	29	31	33	35	37	39	41	43	45	47	49	51	53	55	57	

Setzen wir den Scheidepunkt SP bei Testwert 36 an, so haben wir 9/31 = 29% schulreife Kinder zu Unrecht zurückgestellt, auf der anderen Seite aber keinen ebenso großen Anteil Schulunreifer (33/320 = 10%) unter den Schulreifen erhalten. Leider lassen sich die Anteile durch eine andere Wahl von SP nicht mehr annähern[1]).

Zu 2) Selektionsrate:

- Ist wegen hohen Bedarfs keine Selektion (Selektionsrate = 1) möglich, ist es sinnlos einen hoch validen Test einzusetzen; außer zum Zweck seiner Revalidierung
- Ist Selektionsrate <1, z.B. 0,5 kann der Test die Quote der Geeigneten unter den Eingestellten merklich erhöhen, wenn er eine hohe Validität besitzt.
- Ist sie sehr viel kleiner als 1, z.B. 0,1 dann ist auch ein Test mit geringer Validität von Nutzen, weil man ja nur 10% Pb mit den höchsten Testleistungen benötigt

Zu 3)

- Testauslese ist überflüssig, wenn unter eine Bewerberpopulation nur Geeignete sind.
- Eignungsauslese wird erst mit sinkendem Anteil Geeigneter in der Population zunehmend interessanter und effizienter in dem Maße, in dem auch die Selektionsrate abnimmt.

Ob man einen hoch validen oder einen niedrig validen Test hernimmt, hängt also von der Selektionsrate und dem Anteil der Geeigneten ab.

<u>Modellstudie von RAATZ (1978)</u>

Zeigte, dass es manchmal nicht sinnvoll ist, überhaupt einen Test – auch einen hoch validen – zur Auswahl einzusetzen, da der Anteil an Fehlentscheidungen sich kaum verringerte.

Es ergab sich u.a.

- Wenn beide Arten von Fehlern berücksichtigt werden, dann ist der Einsatz eines Tests nicht sinnvoll, wenn die Selektionsrate sehr hoch oder sehr niedrig, und wenn der Anteil der Geeigneten ebenfalls sehr hoch oder sehr niedrig ist.
- Wenn nur der „institutionelle“ Fehler betrachtet wird, dann ist die Verwendung eines validen Tests anstatt eines weniger validen Auswahlverfahrens nicht notwendig, wenn fast alle Bewerber aufgenommen werden, oder wenn fast alle geeignet sind.