2. TESTTHEORETISCHE GRUNDLAGEN

LIENERT: Unterscheidung von **Haupt- und Nebengütekriterien von Tests**:

Hauptgütekriterien:
* Objektivität
* Reliabilität
* Validität

Nebengütekriterien:
* Normierung
* Ökonomie
* Nützlichkeit

andere Gütekriterien außerdem:
* Zumutbarkeit
* Unverfälschbarkeit der Ergebnisse
* Fairneß
* Skalierung

TESTKENNWERT

= vorschriftsmäßig zu verrechnenede Größe, die die mit dem Test zu messen beabsichtigte Eigenschaft operationalisiert (d.h. eine Variable)

Beispiel: IQ; ist wie alle Testkennwerte theoretisch für unendlich viele TP nach besonderen Verrechnungsvorschriften (je nach Test verschieden) bestimmbar;

Testergebnis = der für eine TP berechnete Testwert, bezogen auf den Testkennwert.

TESTWERT

= ein ganz bestimmtes numerisches Ergebnis dieses Testkennwerts (d.h. die konkrete Realisierung einer Variable)

Merke: ***Testwert*** = ganz allgemein zutreffende Bezeichnung, unabhängig davon, ob es sich um normierte oder nicht normierte Testergebnisse handelt.

Rohscore = Anzahl gelöster Items; auf nicht normierte Testergebnisse beschränkt.

[In der Praxis werden beide Begriffe allerdings synonym verwendet...]

TESTHANDBUCH

= Manual, Handanweisung für den Testleiter, wie Test durchzuführen und zu verrechnen ist (Normtabellen) samt Darstellung des theoretischen Konzepts des Tests

Mag. Margarete PÖKL, 7300178

1. OBJEKTIVITÄT:

Definition:
Unter Objektivität eines Tests versteht man den Grad, in dem die **Ergebnisse eines Tests unabhängig vom Untersucher** sind.

➔ verschiedene TL sollen möglichst zu denselben Ergebnissen kommen (theoretisch, denn in der Praxis im Regelfall KEINE genau gleichen Bedingungen herstellbar, bzw. Übungseffekt)

Man unterscheidet zwischen:

a) Durchführungsobjektivität
b) Auswertungsobjektivität und
c) Interpretationsobjektivität

ad a) **Testleiterunabhängigkeit: (= Durchführungsobjektivität)**

D.h. Testverhalten der TP und Testergebnis sollen **unabhängig sein von zufälligen oder systematischen Verhaltensvariationen des TL**
Für maximal hohe Durchführungsobjektivität -> **Interaktion zwischen TP und TL auf Minimum reduzieren** (Instruktionen schriftlich, standardisierte Untersuchungssituation)

Beeinträchtigung der Objektivität z.B. durch:

- **TL-Erwartungseffekt = Rosenthal-Effekt**
  (-> unwissentliche Verstärkung des Verhaltens der TP meist mittels non-verbaler Kommunikation [Gestik, Körperhaltung, Mimik, Augenbewegungen, Kopfhaltung; Sprache; sozialer Abstand], weil TL eben gewissen Erwartungen an TP hat)

- **Halo-Effekt**
  (TL läßt sich bei Beurteilung z.B. einer Persönlichkeitseigenschaft vom Gesamteindruck oder einem hervorstechenden Merkmal der TP leiten)

  ⇨ **HAWIK**: **geschlechtsspezifischer TL-Effekt** -> Mädchen von Männern getestet schneiden im Verbalteil im Durchschnitt signifikant besser ab als Mädchen von Frauen getestet (bei Buben auch, aber nicht signifikant)

DAHER: in pharmakologischen Experimenten -> **Double-Blind** (d.h. weder VL noch TP weiß, wer zu VG und KG gehört)

Merke: bei **Einzeltests geringere Testleiterunabhängigkeit als bei Gruppentests** -> weil situationsabhängiges Eingehen des TL auf die TP hier unumgänglich ist!

Gruppentests erlauben weit weniger Interaktion zwischen TL und TP, daher **höhere Durchführungsobjektivität**!

Mag. Margarete PÖKL, 7300178

ABHILFE: Computerdiagnostik

➔ hier jede **Interaktion zwischen TL und TP ausgeschaltet** (aber auch hier könnten TL-Effekte auftreten, je nachdem mit welchen Worten und welchem Verhalten dies einweisende Hilfskraft die TP an den PC setzt...)

Besonders problematisch in bezug auf Durchführungsobjektivität ist das **Assessment-Center**

ad b) **Auswertungsobjektivität = Verrechnungssicherheit:**

d.h. Reglementierungen im Manual, wie die einzelnen Testleistungen zu Testwerten zu verrechnen sind, müssen so exakt festgelegt werden, daß **jeder Auswerter zu denselben Ergebnissen** kommt.

abhängig auch vom: -> ***Antwortformat***

Auswertungsobjektivität = hoch:

- wenn TL **nur nach richtig oder falsch zu bewerten** ist (kann bei freiem Antwortformat problematisch werden, wenn Antwort nicht EINDEUTIG richtig oder falsch ist) ; => ist Verrechnung **teilrichtiger Antworten vorgesehen -> Senkung der Auswertungsobjektivität**

- bei **Multiple-Choice-Format** (bzw. festen Antwortformaten)

- bei **Computertests**

- bei **Gruppenverfahren** (weil diese meist nach dem Multiple-Choice-Verfahren gestaltet sind)

Auswertungsobjektivität = geringer:

- bei **projektiven Verfahren** (z.B. bei Familie in Tieren -> Information wird von TL zu TL verschieden beobachtet und ausgewertet)

- bei **freiem Antwortformat** (z.B. im HAWIE-R „Warum muß man Steuern zahlen?“ -> gemeint ist aber nicht „warum“ sondern „wozu“!)

- bei **Individualverfahren** (weil diese oft das freie Antwortformat verwenden)

ad c) **Interpretationsobjektivität = Interpretationseindeutigkeit**:

Ist gegeben, wenn **verschiedene Interpreten** aus denselben Auswertungsergebnissen **dieselben Schlüsse** ziehen

- Ist **bei normierten Tests IMMER** gegeben über die Prozentränge (z.B. PR=95 -> nur 5% der Referenzpopulation erzielen bessere Leistungen)

- **Projektiven Verfahren sind nicht interpretationseindeutig** (sind an tiefenpsychologischen Theorien orientiert -> Interpretation hängt von subjektiver Deutung und Bewertung der Testreaktion durch den TL in bezug auf gesamten Kontext der TP-Umwelt ab

ABER. **Interpretationseindeutigkeit bedeutet NICHT Konsequenzenverbindlichkeit**
(d.h. die Objektivität der diagnostischen Konsequenzen) -> wäre psychologisch nicht verantwortbar (z.B. alle Kinder mit IQ unter 85 in die Sonderschule einweisen...)

[**Kriteriumsorientierte Tests der Pädagogischen Psychologie** sind **interpretationseindeutig UND konsequenzenverbindlich** -> z.B. wer bestimmten Lehrstoff bis zu einem gewissen Grad beherrscht, wird als qualifiziert bezeichnet]

2. Reliabilität:

Definition:
Unter Reliabilität eines Tests versteht man den **Grad der Genauigkeit, mit dem er ein bestimmtes Merkmal mißt, gleichgültig, ob er dieses Merkmal auch zu messen beansprucht.**

Reliabilität = Zuverlässigkeit, formale Exaktheit der Merkmalserfassung; unabhängig davon, was gemessen wird, sollten die unter gleichen Bedingungen gewonnenen Testwerte einer TP gleich sein.
➔ ist **in Diagnostik nur theoretisch möglich** (wegen der Übungs- und Erinnerungseffekte! (daher diverse „andere“ Reliabilitäten...)

a) **Paralleltest-Reliabilität:**

Testwiederholung mit dem ursprünglichen Test ist nicht möglich, dafür aber mit einer **völlig gleichwertigen Nachahmung dieses ursprünglichen Tests**, nämlich einem Paralleltest.

Korrelation zwischen Test und Paralleltest, ermittelt an einer Stichprobe von TP, beschreibt das **Ausmaß, in dem die Relationen der Testwerte der einzelnen TP zueinander konstant** bleiben. (konstante Übungs- und Erinnerungseffekte gehen nicht in die Korrelation ein).

Mag. Margarete PÖKL, 7300178

Geprüft wird aber NICHT die Konstanz der Messung, sondern nur die Konstanz der Ergebnis-Relationen eines Meßinstruments (Tests)

Kritik:

- **große Probleme bei der Erstellung von Paralleltests in der Klassischen Testtheorie**. Korrelation zwischen Test und Paralleltest ist eigentlich nur ein Kriterium dafür, wie gut die Konstruktion des Paralleltests gelungen ist, weniger für die Zuverlässigkeit der Messungen!

- In **probabilistischer Testtheorie** wäre die Konstruktion von Paralleltests einfacher und kontrollierbarer, dort wird Problem der Zuverlässigkeit einer Messung aber **nicht korrelationsstatisch gelöst**...

b) **Retest- Reliabilität (Stabilität):**

Über einen längeren Zeitraum hinweg liefert **die Korrelation zwischen Test und Retest** (Testwiederholung) relevante Informationen über den Test bzw. die mit ihm gemessene Eigenschaft.

=> **Retest-Reliabilität = Stabilität**

ABER: Auch wenn Stabilität relativ hoch ist, müssen Schlüsse für die fernere Zukunft mit Vorsicht gezogen werden!

c) **Innere Konsistenz (r_{tt}):**

= Homogenität im Sinne der Klassischen Testtheorie. D.h. **Items eines Tests** müssen immer innere Konsistenz zeigen, wenn sie **alle der Messung ein- und derselben Eigenschaft dienen.**

Passen die Items nicht zueinander (d.h. **innere Konsistenz = gering**) -> Testergebnis hängt (vor allem bei Parallel- und Re-Testungen) besonders stark von Zufälligkeiten ab **-> Exaktheit der Messung (Reliabilität) ist gering**.

- Bekannteste Methode zur Bestimmung der inneren Konsistenz

  = ***split-half-Methode***:

  Test wird in zwei Hälften geteilt [meist in Items mit gerader und Items mit ungerader Nummer = **odd-even-Methode**] -> Korrelation über die Testwerte beider Teile ausgerechnet. Dann Aufwertung auf gesamte

Testlänge mit der **Spearman-Brown-Formel** -> **Reliabilität des Gesamttests.** [Formeln auf S 38f]

Mag. Margarete PÖKL, 7300178

Abschätzung der Reliabilität (und zwar ihrer unteren Grenze) gibt das **Cronbach-Alpha**

- ***Konsistenzanalyse:***
  Test wird nicht nur in zwei Teile geteilt, sondern in soviele Teile, wie er Items aufweist. Formel dazu = **Kuderman-Richardson-Formel** [Buch S 39]

  Merke: gilt aber **nur für dichotome Items**! Setzt voraus, daß alle Interkorrelationen gleich sind -> wenn das nicht so ist, dann bloß eine Abschätzung der Reliabilität.

MERKE: Ein ***für die Praxis geeignetes Verfahren hat Reliabilitäten um 0,9 oder höher***, ist die Reliabilität niedriger oder nicht angegeben -> diesen Test NICHT verwenden

Standardmeßfehler (SMZ):

mit ihm kann für die TP v mit dem Testwert Xv das **Konfidenzintervall berechnet** werden, **innerhalb dessen ihr wahrer (= meßfehlerbereinigter) Testwert** Tv liegen wird, und zwar **bei festgelegter Irrtumswahrscheinlichkeit Alpha**.
=> Formel dafür auf S 40!

Wozu braucht man den SMZ?

➔ gibt die **Meßungenauigkeit** an! Auch bei hoch reliablen Tests ist mit einer relativ großen Ungenauigkeit zu rechnen!!!

Mag. Margarete PÖKL, 7300178

3. VALIDITÄT:

Definition:
Validität eines Tests gibt den **Grad der Genauigkeit** an, **mit dem der Test das Merkmal, das er messen soll, tatsächlich mißt.**

Merke: Reliabilität ist eine NOTWENDIGE, aber nicht HINREICHENDE Bedingung für einen Test. **Validität** hingegen ist **das wichtigste Gütekriterium** (es ist auch am schwierigsten zu prüfen...)

a) **inhaltliche Validität: (auch Kontentvalidität)**

Liegt vor, wenn der **Test selbst das optimale Kriterium des interessierenden Merkmals** ist.
Synonym verwendet werden dafür auch ***logische Validität*** und ***triviale Validität*** (sind laut Kubinger aber nur Spezialfälle der inhaltlichen Validität!)

Inhaltliche Validität **erreicht man bei einem Test dadurch**, daß man bei **der Konstruktion der einzelnen Items ganz bestimmte definitorisch festgelegte Regeln anwendet**. Nur die Anwendung dieser Regeln führt zu einer richtigen Lösung.

Man erreicht sie am leichtesten **durch ein Experten-Rating** (jedes **Item** wird darauf **geprüft**, ob es **in bezug auf die gegebene operationale Definition** dessen, was der Test messen soll, paßt)

Nicht verwechseln darf man die inhaltliche Validität **mit** der

⇨ **Augenscheinvalidität (face-validity):**

= **der TP ist** augenscheinlich **klar, was der Test zu erfassen** beabsichtigt.

Hohe Augenscheinvalidität kann

- ***von Vorteil sein***:
  wenn TP der Testung Skepsis entgegen bringt (z.B. z.B. Bewerber für Manager-Job lassen sich nur ungern testen, es sei denn mit einem PC-Test, bei dem Geschicklichkeit im Organisieren verlangt wird

- ***von Nachteil sein***:
  vor allem bei Persönlichkeitsfragebogen in der Personalauswahl -> Verfälschbarkeit (z.B. Fragen wie „Ich wache morgens meist frisch und ausgeruht auf.“ werden meist in Richtung sozialer Erwünschtheit beantwortet und nicht ehrlich)

Mag. Margarete PÖKL, 7300178

b) **Konstruktvalidität**:

Ein Test hat Konstruktvalidität, wenn er **gewisse theoretische Vorstellungen erfüllt**.

[***Konstrukt*** = allgemein bekanntes, aber nicht direkt beobachtbares Phänomen, z.B. Intelligenz, Streß, Angst, usw.]

Zur ***Konstruktvalidierung*** herangezogen werden folgende Verfahren:

- **Faktorenanalyse** (angewendet z.B. beim HAWIE):
  Problem = Stichprobenabhängigkeit!

  ➔ **konvergente Validität**:
  bei der Konstruktvalidierung werden konstruktnahe Tests herangezogen
  ⇨ Lädt der Test in denselben Faktoren hoch wie die konstruktnahen Tests = **hohe konvergente Validität [= gut]**

  ➔ **diskriminante Validität**:
  bei der Konstruktvalidierung werden konstruktferne Tests herangezogen
  ⇨ Lädt der Tests in denselben Faktoren niedrig, in denen die konstruktfernen Tests hoch laden = **hohe diskriminante Validität [= gut]**

- **Campbell & Fiske:**

  Nicht nur konstruktnahe und konstruktferne Tests werden zur Validierung herangezogen, sondern alle Tests und alle Methoden. Die Korrelationen bilden die **multi-trait-multi-method-Matrix** (MTMM). In ihr unterscheidet man dann zwischen

  ➔ heteromethodischen Korrelationen
  ➔ monomethodischen Korrelationen
  ➔ heterotrait Korrelationen
  ➔ monotrait Korrelationen

  ⇨ alle monomethodischen-monotrait-Korrelationen drücken die Reliabilität des Tests in bezug auf ein bestimmtes Material aus
  ⇨ alle heteromethodischen-monotrait-Korrelationen drücken die innere Konsistenz (bezogen auf verschiedene Materialien) aus
  ⇨ alle monomethodisch-heterotrait-Korrelationen zwischen den konstruktnahen Tests entsprechen der konvergenten Validität
  ⇨ alle monomethodisch-heterotrait-Korrelationen zwischen den konstruktfernen Tests entsprechen der diskriminativen Validität-

Mag. Margarete PÖKL, 7300178

Welche Bedingungen muß ein valider Test erfüllen (Campbell & Fiske):

1) heteromethodisch-monotrait-Korrelationen müssen wesentlich größer als Null sein (wenn nicht: Methoden messen verschiedene Konstrukte): Nur multimethodisch meßbare Konstrukte sind allgemeingültig (d.h. sie sind interessant für die Persönlichkeitsdiagnostik)

2) heteromethodisch-monotrait-Korrelationen sollten größer sein als die monomethodisch-heterotrait Korrelationen zwischen konstruktfernen Tests (sonst bilden eher die Methoden ein Konstrukt als die Konstrukte selbst)

3) heteromethodisch-monotrait-Korrelationen sollten größer sein als alle heteromethodisch-heterotrait-Korrelationen zwischen konstruktfernen Tests (sonst keine diskriminante Validität)

4) wird in den Punkten 1-3 statt heteromethodisch-monotrait -> mono- bzw. heteromethodisch-heterotrait-Korrelationen zwischen konstruktnahen Tests eingesetzt, so ist konvergente Validität erfüllt.

Andere Arten der ***Konstrukt-Validierung: (nicht-korrelative!)***

- Analyse interindividueller Unterschiede und Unterschiede zwischen Gruppen (vor allem Extremgruppen sind dabei interessant -> **Extremgruppenvalisierung**)

- **Analyse intraindividueller Unterschiede**:
  Test wird mehrmals vorgegeben (Persönlichkeitstest)
  ➔ bei habituellen Persönlichkeitsmerkamel: keine Unterschiede
  ➔ bei situationsabhängigen Verhaltensweisen (z.B. Stimmung): deutliche Unterschiede

- durch **Selbst- bzw. Fremdbeobachtung während des Tests** (vor allem durch Beobachten des Entscheidungs- bzw. Lösungsprozesses) kann untersucht werden, inwieweit das Konstrukt erfaßt wird

c) **Kriteriumsvalidität:**

[= die **einzige Validität, die eine statistische Kennzahl liefert**!
➔ **Kriteriumsvalidität** ist daher **der nur argumentierten inhaltlichen Validität und der Konstruktvalidität überlegen**!!!]

Eine bestimmte als **relevant angesehene Variable (Außenkriterium) wird mit dem Test korreliert** [Problem dabei ist die Auffindung eines solchen Außenkriteriums, bzw. wenn ein solches gefunden ist, daß es nicht 100%ig reliabel sein wird...]

1) **Übereinstimmungsvalidität:**

   bezieht sich auf **die Korrelation mit einem anderen Test, der dasselbe Konstrukt erfaßt** [Aber: wenn der alte Test ohnehin gut mit dem Außenkriterium korreliert, braucht man eh keinen neuen...]

2) Vorhersagegültigkeit = **prognostische Validität**:

   = **Korrelation des Tests mit einem Außenkriterium, das in der Zukunft liegt** (z.B. Prüfungserfolg), d.h. ein bestimmtes später beobachtbares Kriterium soll vorhergesagt werden.

   [**Problem**: viele **Merkmale sind relativ instabil** -> wie weit in die Zukunft soll dann was vorhergesagt werden darüber; Auftraggeber von Gutachten sind aber oft gerade daran interessiert, z.B. Personalchef will, daß der Buchhalter in 30 Jahren nicht Krida macht... ABER: Valide Prognose eines einmaligen Verhaltens, das selten auftritt, ist unmöglich!]

   [**Problem**: **Test zur Personalauswahl können selten prognostisch validiert werden** -> Chefs müßten auch die Durchgefallenen anstellen und beobachten, um Prozentsatz der fälschlich als ungeeignet Prognostizierten in die Validitätsbestimmung einfließen lassen zu können...]

Merke: ***auch wenig-valide Tests sind für Praxis geeignet*** -> Grund: siehe **Taylor-Russel-Tafeln:**

⇨ ist **Selektionsquote niedrig und Grundquote hoch**, so ist auch **bei geringer Validität die Wahrscheinlichkeit einer richtigen Entscheidung hoch**!

- ***Grundquote:***
  Anteil der Geeigneten in einer unausgelesenen Population
- ***Selektionsquote***:
  Anteil der aus einer Population Auszuwählenden

d) ***neues Validierungskonzept:***

Soll ein **Test valide** sein, muß er **in seiner Diagnose mindestens zwischen zwei Gruppen differenzieren**. Ein bewährtes Förderungsprogramm wird eingesetzt

- zeigt bei den Gruppen, für die es gedacht ist Erfolg -> Förderungsprogramm ist valide,
- wenn es keinen Erfolg zeigt, so ist es nicht valide.

4. Normierung:

Definition:

Ein **Test erfüllt das Kriterium Normierung**, wenn für sein Bezugssystem zur Relativierung des individuellen Testergebnisses (= Normen) folgende Bedingungen erfüllt sind:

- die **Normen sind gültig**, d.h. nicht veraltet
- die **Population, für die die Normen gelten, ist definiert**.
- die für die Erstellung der Normen **herangezogene Stichprobe ist repräsentativ**.

a) **Normieren im Sinn von Relativieren:**

Beispiel: **Untertest Allgemeines Wissen aus HAWIE-R:**

besteht aus 24 Items, die nur richtig oder falsch sein können. Testkennwert X = Anzahl der richtigen Antworten (d.h. alle Testwerte liegen zwischen 0-24). Testwerte in der Normstichprobe waren normalverteilt N (0,1) -> jede lineare Transformation ist erlaubt. Standardisierung könnte also sein: z = X - xquer / s
-> positives z = überdurchschnittliche Testleistungen
-> negatives z = unterdurchschnittliche Testleistungen

Wechsler aber legte **folgende Transformation** fest:
W = (X-xquer / s) * 3 + 10
[Fazit: **Wertpunkte haben Mittelwert 10 und Standardabweichung 3]**

Für den IQ legte Wechsler fest:
IQ = (X-xquer / s) * 15 + 100
[-> **IQ-Skala** hat den **Mittelwert 100 und die Standardabweichung 15**]

⇨ unter der Voraussetzung, daß alle Testkenntwerte X normalverteilt waren, ist auch der IQ normalverteilt. **50% der Referenzpopulation fallen in das Intervall 90 kleinergleich IQ kleinergleich 110**

Merke: ***Normal bedeutet nicht gleichzeitig gesund***!

Beispiel: Mensch hat einige kariöse Zähne, das mag statistisch normal sein, aber gesund ist es nicht. Wenn ein Mensch

nie in seinem Leben zum Zahnarzt muß, weil er so gesunde Zähne hat, so ist das dennoch statistisch anormal...

In den vielen Testpublikationen haben sich auch **andere Transformationen** etabliert (was aber nicht heißt, daß die daraus resultierenden Normwerte eine andere oder gar höhere Rechtfertigung hätten.

➔ sie **beruhen alle auf den z-Werten (AUSSER den Stanine-Werten)**

➔ laut Standardnormalverteilungstabelle sind **z-Werte außerhalb -3 kleinergleich z kleinergleich +3 äußerst unwahrscheinlich.**

Gebräuchlichste Werte: (Rest siehe Buch S 56)

	Formel	***Mittelwert***	***Standardabweichung***
1) z-Werte:	-3 <= z <= +3	0	1
2) Wertpunkte:	W = z * 3 + 10	10	3
3) Intelligenzquotient:	IQ = z * 15 + 100	100	15
4) Z-Werte (=SW): Standardwerte	Z = z * 10 + 100	100	10
5) T-Werte:	T = z * 10 + 50	50	10

Relativierung des Testwerts in bezug auf die Referenzpopulation gelingt **am anschaulichsten**, wenn man die Testwerte umrechnet in

Prozentrangwerte:

- Ein PR gibt den **relativen Anteil von Personen** in der Referenzpopulation an, die **denselben oder einen niedrigeren Testwert** haben.

- PR werden **auch aus der Standardnormalverteilung abgeleitet**, sind die **einzigen Normwerte, die NICHT aus einer Lineartransformation stammen**.

- PR sind **auch bei nicht-normalverteilten Testwerten eruierbar** !!!, weil sie unmittelbar aus der Häufigkeitsfunktion der Normierungsstichprobe empirisch bestimmt werden.

- **Diagnostisch könnte man mit PR-Werten das Auslangen** finden, für wissenschaftliche Evaluationen aller Art braucht man jedoch die normalverteilten Normwerte.

Graphische Darstellung aller Transformationen im Buch S 57 !!!

Aus den **Normierungstabellen im Testhandbuch** kann man für jeden Rohscore den entsprechenden Normwert ablesen (aufpassen, daß man die richtige Tabelle erwischt -> z.B. Altersnorm, Geschlechtsnorm, usw. was man halt je nach Fragestellung braucht)

Es gibt auch **Nomogramme** (z.B. im WMT) -> man erspart sich damit umfangreiche Tabellen, weil z.B. verschiedene Altersgruppen hier simultan dargestellt werden können. Mit ihnen ist ein sehr anschaulicher, unmittelbarer Vergleich zwischen verschiedenen Gruppennormen möglich

Merke: In der Praxis werden **oft voreilige Schlußfolgerungen** darüber gezogen, was die ***Unterschiede zweier Testwerte*** betrifft

⇨ für jeden der beiden Testwerte ein **Konfidenzintervall** ausrechnen und schauen, ob sie sich überschneiden oder nicht:

- wenn ***Überschneidung -> kein signifikanter Unterschied***
- wenn ***keine Überschneidung -> signifikanter Unterschied***

b) **Normierung im Sinn von Repräsentativerhebung:**

Es ist schwierig ***Repräsentativität für die Normierungsstichprobe*** zu erreichen.

- **Stichprobe** muß **ausreichend groß** sein; ABER: mangelnde Repräsentativität kann nicht durch Größe kompensiert werden!

- Stichprobe muß auch **repräsentativ für die Population** sein, aus der sie genommen wurde und über die mittels des Tests später etwas ausgesagt werden soll!

Merke: Auch **Normen sind einem Wandel unterworfen**... (vgl. Studien, die bezeugen, daß der durchschnittliche IQ, gemessen am selben Test innerhalb von ca. 20 Jahren um bis zu 25 Punkte gestiegen ist.

Merke: Man sollte **in der Praxis KEINEN Test einsetzen, der älter als 10 Jahre** ist, wenn er nicht zwischenzeitlich überarbeitet, auf seine Güte kontrolliert und neunormiert wurde! (-> wegen der testtheoretischen Entwicklungen, der differentialdiagnostischen Erkenntnisse und der gesellschaftlichen Veränderungen, die in dieser Zeit stattgefunden haben!)

Merke: Laut **AID** gibt es **keinen Unterschied hinsichtlich der Intelligenz zwischen Österreich und Deutschland**. Es gibt jedoch Unterschiede

in bezug auf einzelne Persönlichkeitseigenschaften zwischen BRDlern und DDRlern (z.B. Fähigkeit zum Erfolg bei DDRlern laut CPI weniger deutlich ausgeprägt!)

Merke: Es ist **relativ schwierig**, genügend **Erwachsene für Normierungsstichprobe** zusammenzubringen, leichter geht das mit Schulkindern;
ABER: die Population der Schulkinder ist nicht ident mit der Population aller schulpflichtigen Kinder!

Anstatt Zufallsstichprobe im klassischen Sinn zu erheben, genügt auch eine sogenannte „***Klumpenerhebung***", z.B. einzelne Schulen und deren Schüler (ABER: Klumpen müssen zufällig gewählt werden, am besten nach regional festgelegten Quoten!)

Zusätzlich sollte in bezug auf wesentliche Variablen der Aspekt der ***Quotenstichprobe*** berücksichtigt werden, d.h. verschiedene Schultypen, Altersstufen, Geschlechter sollten in der Stichprobe den relativen Anteilen in der Population entsprechen.

Beispiel: **AID** -> Durchschnittliche Testwerte weichen pro Sozialschicht stark voneinander ab -> im AID ergab sich eine deutliche Überrepräsentierung der Stadtkinder gegenüber der Landkinder (zum Glück unterscheiden sich diese beiden Gruppen jedoch nicht signifikant in ihren Testwerten!)

c) **Kriteriumsorientierte Diagnostik:**

In vielen Fällen ist eine normorientierte Diagnostik nicht zweckmäßig ; Alternative dazu = kriteriumsorientierte Diagnostik:

Definition:

Kriteriumsorientierte Diagnostik bedeutet, daß **andere Vergleichsmaßstäbe zur Interpretation des Testwerts herangezogen werden als die Testwertverteilung in der Referenzpopulation** (z.B. ein absolut oder individuumsbezogener, relativ gewählter Ziel-Testwert).

Wird vor allem **in der Pädagogischen Diagnostik bzw. förderungsorientierter Diagnostik angewandt**.

Beispiel: in Eignungsdiagnostik nützt ein hoher Prozentrangwert nichts, wenn damit nicht ein bestimmtes Kriterium erfüllt wird -> Sekretärin mit PR von 98% ist wegen vieler Rechtschreibfehler in kurzem Diktat trotz des hohen PR ungeeignet...
ABER: z.B. eine Teilleistungsstörung wäre ohne Verwendung von Normen nicht identifizierbar!

Fazit: normorientierte und kriteriumsorientierte Diagnostik müssen einander ergänzen bei Entscheidung, WELCHE Normen die relevanten sind:
Beispiel: Kind aus unterer Sozialschicht -> Orientierung an Gesamtnormen nützt wenig, besser Orientierung an

schichtspezifischen Normen: Ist Kind allgemein unterdurchschnittlich, schichtspezifisch aber überdurchschnittlich, so ist die Prognose für dieses Kind sicherlich gut.

5. ***SKALIERUNG:***

Definition:

Test erfüllt das Gütekriterium Skalierung, wenn **die laut Verrechnungsvorschriften resultierenden Testwerte die empirischen Verhaltensrelationen adäquat abbilden**. D.h. wenn sie diese messen.

[***Messen:*** ***=*** ist die Bestimmung des Ausprägungsgrades einer Eigenschaft eines Objekts, erfolgt durch Zuordnung von Zahlen zu diesem Objekt. Zuordnung ist aber nur dann eine Messung, wenn die numerischen Beziehungen zwischen Meßwerten empirische Beziehungen zwischen Meßobjekten ausdrücken.]

a) **Methoden der Skalierung:**

Klassische Testtheorie:

Ein Test ist gut, wenn:

1) **innere Konsistenz möglichst hoch ist** (d.h. alle Items messen dasselbe)

2) Faktorenanalyse (d.h. die **Items sollen auf einem einzigen Faktor hoch laden**)

3) **Interkorrelationen aller Items sollen nahezu 1 betragen**

   Merke: **FA geht nur bei nicht dichotomen Items** -> wird die FA auf dichotome Variablen angewendet, so führt die zu artifiziellen Faktoren!

4) **Schwierigkeiten der Items** (= bei dichotomen Items eines Leistungstests sind das die **relativen Lösungshäufigkeiten**, mit denen sie in der Normierungsstichprobe gelöst wurden) sollen **gleichmäßig innerhalb eines Intervalls von 0,05 - 0,095 liegen**

5) **Trennschärfenindizes der Items** (= Korrelationen des Testwerts pro Item mit dem Gesamtscore) sollen **sehr hohe Werte annehmen**

⇨ **Kritik an den Methoden der Klassischen Testtheorie**:

Alle sind **extrem stichprobenabhängig**!!! (z.B. Trennschärfe, Itemschwierigkeit sind stets von Fähigkeit der Stichprobe abhängig -> bei schlechter Stichprobe ganz

andere Werte als bei guter!) -> mit geeigneter Stichprobenauswahl kann praktisch jeder beliebige Wert erreicht werden = Manipulation!

Probabilistische Testtheorie:

Sie **kann prüfen**, ob die gegebenen Verrechnungsvorschriften eines Tests zu Testwerten führen, die verhaltensadäquate Relationen wiedergeben. Mit ihr kann man feststellen, **ob ein Test verrechnungsfair** ist.

Einfachster Verrechnungsmodus = Anzahl gelöster Items. Unabhängig davon, welche Items eine TP gelöst hat und welche nicht, zählen nur die Treffer. Dieser **Verrechnungsmodus ist fair, weil das Rasch-Modell gilt**! Tests, für die das RM nicht gilt, sind nicht verrechnungsfair.

Rasch-Modells:

- es geht **nur um eine Eigenschaftsdimension**, die gemessen werden soll

- es gilt **stochastische Unabhängigkeit** (d.h. ob TP Item löst oder nicht, hängt nur von ihrer Fähigkeit ab und von der Schwierigkeit des Items, nicht davon welche anderen Items sie schon gelöst hat!)
  RM beschreibt die Wahrscheinlichkeit, daß eine TP v ein Item i löst in Abhängigkeit vom Personenparameter (= die wahre Fähigkeit der Tpv) und des Itemparameters (= wahre Schwierigkeit von i)
  RM ist stichprobenunabhängig!
- RM hat **spezifische Objektivität**, d.h. Unterschied in den Fähigkeiten zweier Personen kann unabhängig davon bestimmt werden, welche Items des Tests dazu herangezogen werden und Unterschied zwischen Schwierigkeiten zweier Items kann unabhängig davon bestimmt werden, welche TP damit getestet wurden

- **Geltung des RM kann mittels Modelltests überprüft werden**, z.B.

➔ ***graphisch***: 2 Personenstichproben -> Vergleich der geschätzten Itemparameter. Parameterschätzungen pro Item werden in einem rechtwinkeligen Koordinatensystem eingetragen, Gerade im Winkel von 45° durch den Nullpunkt: liegen Punkte auf der Geraden -> RM gilt.

➔ mittels des ***Likelihood-Quotienten-Tests (***nicht modell-konforme Items werden ausgeschieden)

⇨ **Für folgende Tests gilt das RM:**

WMT, MTP, AID, 3DW, WST, META, RIS, SYL.

-> diese Tests sind verrechnungsfair

⇨ (**Beispiel für einen nicht-verrechnungsfairen Test = HAWIK:**

z.B. Subtest „Allgemeines Wissen“ -> Frage: wieviele Menschen gibt es auf der Welt? Dieses Item fällt Buben relativ leichter als Mädchen -> es mißt nicht eindimensional... Hätte in der Entwicklung des Tests überprüft werden sollen und ein solches Item hätte herausgenommen werden müssen! Beim Subtest „Gemeinsamkeitenfinden“ gibt es auch Punkte für teilrichtige Antworten -> Verrechnungsmodus ist komplizierter, daher wären noch strengere Voraussetzungen und Modellansprüche an die Items zu stellen!

Besonderer Verrechnungsmodus = Bestimmung des IQ:

IQ = **durchschnittlicher Testwert in den Untertests einer Testbatterie**. Es ist egal, in welchem Subtest hohe und in welchem niedrige Leistungen erzielt wurden. [Was aber an sich problematisch ist... Leistungshoch im Puzzlelegen kann ein Tief im Logischen Denken wohl nicht wirklich kompensieren...] Um die Fairneß dieser Verrechnungsart zu prüfen, braucht man das Poisson-Modell von Rasch

b) **Skalierung und Reliabilität:**

In der **Probabilistischen Testtheorie** ist es auch möglich, **Meßgenauigkeit eines Tests mit anderen Maßen als mit Korrelationen zu bestimmen.**
Bei allen Modellen erfolgt die Schätzung der Parameter nach der **Maximum-Likelihood Methode** -> Erkenntnisse aus der Statistischen Mathematik sind daher anwendbar:

Wenn aus unendlich vielen Stichproben einer Population immer wieder derselbe Parameter geschätzt werden soll, so sind diese Schätzungen nicht ident, sondern unterliegen einer gewissen Streuung. Die Varianz dieser Schätzungen (z.B. des Personenparameters einer Tpv) ist dann gleich dem Kehrwert der Informationsfunktion. (Formel Buch S 76)

D.h. die **Genauigkeit des Testergebnisses pro Person ist unabhängig von den sonst noch getesteten Personen, wenn in die Formel anstatt der Parameter die Parameterschätzungen eingesetzt werden.**

Merke:

Was die probabilistische Testtheorie im Vergleich zur klassischen nicht braucht:

- Die probabilistische Testtheorie kommt **grundsätzlich ohne das Konzept der Reliabilität** aus. Sie braucht auch keine innere Konsistenz, weil die Items ja rasch-homogen sind und somit eindimensional, bzw. alle dasselbe messen

- die probabilistische Testtheorie braucht auch **keine Paralleltests** zur Bestimmung der Genauigkeit eines Testergebnisses;

- die **Erstellung von Paralleltests ist unproblematisch** -> man braucht nur einen genügend großen Itempool homogener Items -> jede Teilmenge daraus ist wieder rasch-homogen -> so kann man ganz leicht einen Paralleltest herstellen.

c) **Skalierung und Normierung:**

Problem der Übertragung von Tests in eine andere Kultur:

⇨ **Fehlen von signifikanten Mittelwertsunterschieden kann nicht die Frage klären, ob ein Test innerhalb verschiedener Kulturen dasselbe mißt**

⇨

➔ Kulturunterschied könnte sich ja nicht auf die Gesamttestleistung, sondern nur auf einzelne Items auswirken -> hier spricht Gütekriterium Skalierung gegen die Anwendung eines solchen Tests in einer anderen Kultur als in der der Normierungsstichprobe zugrundeliegenden

Beispiel: **HAWIE-R** Untertest „Allgemeines Wissen“ Item: Wer wählt bei uns den Bundeskanzler?“

6. ÖKONOMIE:

Definition:
Ein Test erfüllt das Gütekriterium Ökonomie, wenn er, **gemessen am diagnostischen Informationsgewinn, relativ wenig Ressourcen** (Zeit und Geld) beansprucht.

➔ ***Vorteil der Probabilistsische Testtheorie*** hierbei:
ermöglicht mit Hilfe des adaptiven Testens, daß ein Test trotz geringer Testlänge genau mißt!!!

➔ [Klassische Testtheorie dagegen: höhere Meßgenauigkeit bzw. kleiner Standardmeßfehler erreicht man nur dadurch, daß man Test verlängert!]

a) **adaptives Testen:**

Nachteile der konventionellen Tests:

1) Test soll großen Leistungs- und Altersbereich umfassen -> man braucht dazu **viele Items von sehr leicht bis sehr schwierig**

   Beispiel: HAWIE-R „Allgemeines Wissen“:
   Wieviele Monate hat ein Jahr? - Wie weit ist es von der Erde bis zum Mond?).

   Test wird dadurch recht lang -> hohe Leistungsbereitschaft der TP ist erforderlich;

   ABER: **pro TP sind nicht genug Items informativ**, d.h. TL weiß oft schon im vorhinein, ob TP ein Item lösen wird oder nicht, daher wenig **Informationsgewinn**. Ein solcher **nur bei Items**, von denen es dem TL ungefähr **gleichwahrscheinlich** erscheint, daß **TP sie löst oder nicht** -> das sind aber nur wenige, daher: die Vorgabe von für die TP zu leichten oder zu schwierigen Items = unökonomisch!

   Abhilfe: z.B. altersspezifische Einstiege für die nach Schwierigkeiten geordneten Items

2) **Für Messungen in Extremleistungsbereichen** gibt es, wenn man hier zwischen zwei TP genau differenzieren möchte **trotz großer Testlänge nur wenige Items**

   Beispiel: HAWIK-R „Allgemeines Wissen“: im unteren Leistungsbereich zwischen allen 6-Jährigen differenzieren nur 4 - 5 Items, alle anderen sind zu schwierig). D.h. man kann nur zwischen ganz

wenigen Leistungskategorien unterscheiden, dabei drastische Meßfehler durch Zufallseinflüsse.

3) Darbietung der **Items in der Reihenfolge ihrer Schwierigkeiten** kann **motivationsbeeinträchtigend** sein. Gibt **es keine leistungsentsprechenden Testeinstiege:**

- zu leichte Items am Anfang = **demotivierend**
- zu schwierige Items am Schluß = **frustrierend** (wurscht ob Abbruchskriterium oder nicht)

Beispiel: HAWIE-R und HAWIK-R -> wie soll TP zu neuem Subtest motiviert werden, wenn sie bei jeden vorhergehenden durch Leistungsversagen beendet hat...)

4) Großer Standardmeßfehler von Untertests mit wenig informativen Items -> **Interpretation des Leistungsprofils** (= Identifizierung von Hochs und Tiefs) ist **kaum möglich.**

5) Weil nur wenige Items für eine bestimmte TP informativ sind -> **anstatt dichotom wird oft mehrkategoriell verrechnet** (z.B. durch Berücksichtigung von teilrichtigen Antworten, Vergabe von Zeitpunkten) **-> Problematik solcher Verrechnungen.**

Daher: ***Adaptives Testen***

⇨ ***Grundidee***:

TP werden **nicht alle Items** vorgegeben, sondern **nur die ihrer individuellen Fähigkeit adäquaten** (d.h. Itemauswahl wird dem jeweiligen Leistungsvermögen angepaßt); ABER: das geht **NUR mit Hilfe der probabilistischen Testtheorie**!!!

⇨ ***zwei Strategien des adaptiven Testens:***

1) **tailored-Testing**
(geht aber nur bei Tests, die nach richtig/falsch bewertet werden!)

2) **branched-Testing**
(geht aber nur bei Tests, die nach richtig/falsch bewertet werden!)

ad 1) **Tailored - Testing:**

- Pro TP wird **mit einem durchschnittlich schweren Item begonnen** (Voraussetzung: Itemschwierigkeiten müssen bekannt sein; großer Itempool muß vorhanden sein).

- Löst TP dieses Item -> **schwierigstes (leichtestes) Item wird geboten**.

- Je nach dem Abschneiden dabei wird **nun das Item ausgewählt**, von dem zu erwarten ist, daß es eine **vorläufige Schätzung des Fähigkeitsparameters erlaubt** (ist der Fall, wenn TP zumindest ein Item gelöst und eines nicht gelöst hat).

- Ist diese **Schätzung gelungen** -> **Auswahl jener Items, die in bezug auf ihre Schwierigkeit mit dem Leistungsfähigkeitsniveau der TP übereinstimmen** (= solche Items, die maximale Information leisten)

➔ **relativ bald ist eine ungefähre Schätzung des Personenparameters gegeben**, sie wird mit jedem weiteren Item genauer (d.h. Schätzung des Personenparameters weicht mit Fortdauer des Tests immer weniger vom wahren Wert des Personenparameters ab)

⇨ ***Wann ist die Testvorgabe zu beenden?***

wenn die **Differenz der Schätzungen ein und desselben Personenparameters bei zwei aufeinanderfolgenden Items nicht mehr größer** ist als ein gewisser geringer, vorher festgelegter Betrag

⇨ ***Wie groß muß der Itempool sein?***

theoretische Erfahrungen ergaben, daß **ca. 60-70 Items ausreichen**, meist ist nach ca. 15 Items hinreichende Genauigkeit des Fähigkeitsparameters erreicht.
[ist **Itempool zu klein** -> zu wenig informative Items -> Kriterium, daß bei zwei aufeinanderfolgenden Items die Schätzungen desselben Personenparameters nur mehr geringfügig voneinander abweichen, kann nicht erfüllt werden]

⇨ ***Wie sind die Testleistungen von 2 TP vergleichbar***, denen völlig andere Items vorgegeben wurden?

Anzahl der gelösten Items ist als Testwert ungeeignet [bei einer TP wären das z.B. 15 leichte Items, bei einer anderen 15 schwierige Items **-> empirische Verhaltensrelationen würden nicht adäquat abgebildet** werden]. In **Probabilistischer Testtheorie** ist es aber möglich, den **unbekannten Personenparameter aus der jeweiligen Modellgleichung**, also unter Berücksichtigung der getroffenen Itemauswahl, zu **schätzen, wenn die Itemparameter bekannt sind..**

Vorteile des adaptiven Testens:

1) bei gleicher Testlänge kann gegenüber einem konventionellen Test eine **wesentlich größere Meßgenauigkeit** erzielt werden (=> ein halb so langer adaptiver Test reicht aus, um ebenso genau zu messen wie ein konventioneller Test!)

2) beim adaptiven Testen wird **in allen Leistungsbereichen gleich genau gemessen** (konventioneller Test dagegen enthält viele mittelschwere Items -> mißt in diesem Bereich genau, nicht aber in den Extremleistungsbereichen, da hier jeweils zu wenige informative Items vorhanden)

3) beim adaptiven Testen **keine motivationalen Beeinträchtigungen** zu erwarten -> Itemauswahl für die einzelne TP erfolgt so, daß sie die Items abwechselnd löst und nicht löst -> keine Frustration, keine Demotivation

4) beim adaptiven Testen **höhere Meßgenauigkeit pro** Untertest -> Leistungsprofil ist daher interpretierbar

5) beim adaptiven Testen werden höchst informative Items verwendet **-> keine mehrkategorielle Verrechnung nötig**

Nachteile des adaptiven Testens:

1) Für adaptives Testen braucht man größere Anzahl von Items als für konventionellen Test -> **Testkonstruktion ist aufwendiger**; testtheoretische Analysen im Sinne der Probabilistischen Testtheorie sind notwendig

2) **Realisierung des tailored-testings ist gebunden an Computerdiagnostik** -> Schätzung der Personenparameter ist nur mittels des Computers möglich, weil sie nach Bearbeitung JEDES Items erfolgen muß, damit man das nächste geeignete Item auswählen kann.

DAHER: **derzeit funktioniert NUR der Test Syllogismen nach diesem Prinzip**!

Merke: * **Bei uns** basieren die **adaptiven Tests auf dem Rasch-Modell**,
* in **Amerika** dagegen auf dem **Birnbaum-Modell** (macht aber wesentlich mehr Probleme bei der Schätzung der Parameter, daher Rasch-Modell vorzuziehen)

Merke: **Verwendung ausschließlich informativer Items** (wie beim adaptiven Testen) erlaubt **ökonomisches Testen und trotzdem höchste Meßgenauigkeit**;
ABER: bei **Matrizentests** werden **durch adaptives Testen die Itembearbeitungszeiten auf das Doppelte erhöht**

➔ d.h. Verkürzung der Testlänge bedeutet nicht notwendigerweise auch Verkürzung der Testbearbeitungszeiten!!!

⇨ ***Wieso muß man beim adaptiven Testen mit einer Verlängerung der Itembearbeitungszeiten rechnen?***

TP brauchen zum Aufwärmen leichtere Items, erwarten Items nach der Schwierigkeit geordnet -> werden ihnen jetzt ziemlich bald schwierige Items vorgegeben bzw. kommen zwischendurch immer wieder auch leichtere Items vor -> TP ist noch nicht richtig eingestimmt bzw. irritiert -> Reaktionszeit verlängert.

ad 2) **Branched-Testing:**

- **Nach Bearbeitung eines einzelnen Items** wird **nicht das aktuell informativste Item als nächstes vorgegeben**, sondern
- **Items in Gruppen zusammengefaßt** -> je nach Leistungsgüte wird **in festverzweigter Weise auf eine bestimmte Gruppe verwiesen**

Beispiel: AID „Alltagswissen“:

Begonnen wird, je nach Alter des Kindes mit einer bestimmten Itemgruppe aus 5 Items -> löst Kind nur 1 Item, war Gruppe zu schwer, daher zu leichterer Gruppe
➔ löst Kind mindestens 4 Items, war Gruppe zu leicht, daher zu schwierigerer Gruppe
➔ löst Kind 2-3 Items, paßt Gruppe, nächste Aufgabengruppe mit gleicher Schwierigkeit
Insgesamt: Vorgabe von 3 Gruppe à 5 Items, d.s. 15 Items
Bei Screening-Verfahren zu einer grob klassifizierenden Diagnose genügen auch 2 Aufgabengruppen

⇨ Damit **Testleistungen der TP** von Personen mit unterschiedlich bearbeiteten Items **vergleichbar** werden, **muß Test einem Modell der Probabilistischen Testtheorie entsprechen.**

Besondere technische Fragen des Branched-Testing:

1) **wieviele Startgruppen?**
➔ Meßgenauigkeit wird nicht größer, wenn man möglichst viele Startgruppen hat, daher sind auch 3 Gruppen bereits ausreichend

2) **wieviele Verzweigungsschritte? wieviele Verzweigungsmöglichkeiten?**
➔ Dreifachverzweigung (mit weniger Schritten) ist besser als Zweifachverzweigung (mit mehr Schritten)

3) **wieviele Items pro Itemgruppe müssen gewählt werden?**
➔ besser ist mehr Verzweigungsschritte als mehr Items

4) **Was passiert, wenn man die Startgruppe inadäquat wählt?**
   ➔ keine Auswirkungen; kritisch ist nur der Extremfall, daß Personenparameter wegen ausschließlich gelöster oder nicht gelöster Items nicht zu schätzen ist (aber dafür gibts auch Lösung, nämlich kurvilineare Extrapolation).

BBT = pyramidales Testen:

- ist ein **besonderes Vorgabeschema des branched-Testings**
- anstatt einer Gruppe von Items wird (wie beim tailored-Testing) **nach jedem einzelnen Item je nach Leistungsgüte das nächste ausgewählt, aber in festverzweigter, vorbestimmter Weise** (= Unterschied zum tailored-Testing)

b) **Ökonomie und Aufwandsminimierung:**

Psychologisches Diagnostizieren verfehlt seinen Zweck, wenn:

1) eine **nicht ausreichend ausgebildete Person als TL** fungiert
2) auf den **Einsatz von Tests verzichtet** wird, wenn sie **wertvolle diagnostische Informationen liefern** würden
3) **Test** verwendet werden, **anstatt eine systematische Verhaltensbeobachtung** vorzunehmen, die erst die relevante diagnostische Information bringen würde
4) der **TL nur solche Tests einsetzt, mit denen er seit Jahren vertraut** ist, die aber **weniger diagnostische Informationen** liefern
5) **nur Tests** eingesetzt werden, die **nach dem Krankenkassenvertrag abgerechnet** werden können, obwohl andere den diagnostischen Informationsgewinn steigern könnten
6) die **Testwertung primär danach** erfolgt, daß Testvorgabe oder zumindest die **Auswertung mittels PCs möglich** ist.

7. ***NÜTZLICHKEIT:***

Definition:

Ein Test ist dann nützlich, wenn er ein **Persönlichkeitsmerkmal mißt, für dessen Untersuchung ein praktisches Bedürfnis besteht**.
(d.h. ein Test ist dann nützlich, wenn er durch keinen anderen ersetzt werden könnte)

⇨ ***Wann ist ein Test von Nutzen?***
➔ wenn es innerhalb einer bestimmten diagnostischen Fragestellung notwendig ist, diesen Test einzusetzen.

In der **Selektionsdiagnostik** gibt es **Verfahren, mit denen es möglich ist, denjenigen optimalen Trennscore aller möglichen Testwerte zu bestimmen**, bis zu dem Kandidaten als nicht geeignet und ab dem Kandidaten als geeignet zu bezeichnen sind.

Man unterscheidet ***4 Fälle:***

a) **richtige positive Diagnose:**
(z.B. Feststellung einer Cerebralschädigung, die vorliegt)
b) **falsche positive Diagnose:**
(z.B. Feststellung einer Cerebralschädigung, die nicht vorliegt)
c) **richtige negative Diagnose:**
(z.B. Feststellung einer Nicht-Cerebralschädigung und der Patient hat auch keine solche)
d) **falsche negative Diagnose**
(z.B. Feststellung einer Nicht-Cerebralschädigung, obwohl eine Cerebralschädigung vorliegt)

➔ je nach Nutzenfunktion kann ein und derselbe Test für ein und dieselbe Fragestellung nützlich sein oder nicht.

zwei relevante Nutzenfunktionen:

a) **Test wird als Screening-Verfahren eingesetzt** -> grob klassifizierende Vorauswahl mit dem Ziel, beim geringsten Verdacht den befürchteten Befund positiv zu diagnostizieren

b) Mit dem Test wird **keinesfalls voreilig ein positiver Befund diagnostiziert** -> Ziel = Etikettieren auf Dauer zu vermeiden.

Merke:

- relativer Anteil, mit dem **bei einem gegebenen positiven Zustand die Diagnose richtig** ist, wird in der Medizin ***SENSITIVITÄT*** genannt
- relativer Anteil, mit dem **bei negativem Zustand die Diagnose richtig** ist, wird in der Medizin ***SPEZIFITÄT*** genannt.

8. ***ZUMUTBARKEIT:***

Definition:

Zumutbarkeit ist das **Ausmaß, in dem ein Test** (in bezug zu dem aus der Anwendung des Verfahrens resultierenden Nutzens) **die TP in zeitlicher, psychischer und körperlicher Hinsicht beansprucht.**

Das ist erst seit einigen Jahren in der Diagnostik ein Thema.
Vieles von dem, was traditioneller Zustand ist, ist den TP nur mit Vorbehalt zumutbar

⇨ ***Was ist eigentlich unzumutbar?***

1) Tests mit einer **Dauer von 4 oder mehr Stunden in einer einzigen Sitzung** (**Beispiel**: **WIT**)

2) Persönlichkeitsfragebogen mit **Fragen zum Intimbereich** (Beispiel: MMPI)

3) **Persönlichkeitsfragebogen mit dichotomem Antwortformat** (= forced-choice Format) ohne die Möglichkeit einer neutralen Antwort oder des Nuancierens (Beispiel: MBTI)

4) **projektive Verfahren ohne jede Augenscheinvalidität** (z.B. Rohrschach-Form-Deute-Verfahren)

9. ***UNVERFÄLSCHBARKEIT:***

Definition:

Unverfälschbarkeit eines Tests ist das **Ausmaß, in dem ein Test die individuelle Kontrolle über Art und Inhalt der verlangten bzw. gelieferten Informationen ermöglicht.**

➔ besser wäre allerdings, das Gegenteil dieser Definition als Gütekriterium festzulegen...

- **Leistungstests haben normalerweise eine hohe Augenscheinvalidität**, und das ist auch **wünschenswert**; grundsätzlich ist es möglich, daß eine TP absichtlich schlechte Leistungen erbringt, das ist aber eher unrealistisch

- **Persönlichkeitstest mit hoher Augenscheinvalidität jedoch** bergen die **Gefahr**, daß die TP sozial erwünscht und nicht ehrlich antwortet! (Ausnahme: im klinischen Bereich, denn dort erwartet sich TP ja bei genügend großem Leidensdruck Hilfe von der Diagnostik)

[Hier steht allerlei, was zur Diagnostik II gehört -> im nächsten Semester hier einfügen!]

Merke:

⇨ ***Beispiele, wo TP in Leistungstests absichtlich schlecht abschneiden***:

➔ Tests im Rahmen von militärischen Untersuchungen
➔ Tests im Auftrag von Versicherungsgesellschaften

Hier könnte es dazu kommen, daß sich **TP aufgrund eines schlechten Abschneidens Vorteile** bzw. die Verhinderung von Nachteilen erwartet

10. ***FAIRNEß:***

Definition:
Fairneß eines Tests ist das **Ausmaß einer systematischen Diskriminierung bestimmter Tpn, z.B. aufgrund ihrer ethischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit.**

Jahrzehntelanges Bemühen der Diagnostik um

CULTURE-FAIR-TESTS:

zeigen wie wichtig es ist, beim Diagnostizieren kulturbedingte Handicaps von Tpn zu vermeiden!

⇨ entsprechende Testkonzepte kommen von der **Materialgestaltung her OHNE Sprache** aus (sowohl für Verständnis der Instruktion noch zur Lösung der Items bedarf die TP unmittelbar der Sprache):

- solche Tests weisen „**Sprach-Fairneß** auf; sind fair, weil der Sprachstil belanglos ist

- **ABER: sie sind NICHT sprachunabhängig**, d.h. die Lösung erfolgt in der Regel durch stilles Verbalisieren

Beispiel: Matrizentests
(z.B. WMT [sprachfreie Instruktion ist zwar nicht vorgesehen, wäre aber leicht möglich!], SPM [kommt nicht ohne sprachfreie Instruktion aus]

1) **ethnisch bedingte Probleme der Fairneß** (siehe unter Normierung, dort steht:

- **Kulturunterschied** wirkt sich **meist nicht auf die Gesamttestleistung aus, sondern nur auf einzelne Items** -> Gütekriterium Skalierung spricht sich **gegen die Anwendung eines solchen Tests in einer anderen Kultur als in der der Normierungsstichprobe zugrundeliegenden**:

  Beispiel: HAWIE-R Untertest „Allgemeines Wissen“ Items: Wer wählt bei uns den Bundeskanzler?“ und „Wieviele Einwohner hat Deutschland?“

- **Laut AID gibt es keinen Unterschied hinsichtlich der Intelligenz zwischen Österreich und Deutschland**. Es gibt jedoch Unterschiede in bezug auf einzelne Persönlichkeitseigenschaften zwischen BRDlern und DDRlern (z.B. Fähigkeit zum Erfolg bei DDRlern laut CPI weniger deutlich ausgeprägt!)“

2) **milieubedingte Unterschiede**:

Beispiel: AID

Hier gibt es **keine schichtspezifischen Normen, ABER:** aufgrund **signifikanter Mittelwertsunterschiede** wird im Handbuch darauf hingewiesen, daß Testleistungen eines Kindes zu relativieren sind
➔ allgemein eher unterdurchschnittlich leistungsfähiges Kind kann schichtspezifisch durchaus Überdurchschnittliches leisten, was eine gute Prognose für sein Leistungspotential bedeutet

3) **geschlechtsspezifische Unterschiede:**

Beispiel: Schlauchfiguren

Hier gibt es **eigene Normen für Frauen**
➔ Frauen schneiden bei solchen Aufgaben generell schlechter ab als Männer

4) **Einfluß der Testerfahrung einer TP**:

Manche Tests bevorzugen in hohem Grad **Testroutinees**, viele Tests benachteiligen TP ohne jegliche Testerfahrung. Es geht hier aber weniger um Lern- und Übungseffekte, sondern eher um Gewöhnung und Vertrautheit mit psychologischen Testitems.
➔ Tests sollten daher auch hinsichtlich ihrer Erfahrungsunabhängigkeit überprüft werden!

Beispiel: Kinder mit Testerfahrung in einem anderen Intelligenztest (z.B. AID) **schneiden im HAWIK besser** ab; ABER Kinder mit Testerfahrung schneiden im AID nicht besser ab -> AID ist fairer als HAWIK!

Computerdiagnostik ist, was Erfahrungsunabhängigkeit betrifft, in typischer

Weise problematisch (siehe dort!)

Merke:

In letzter Zeit gibt es **Bestrebungen, Testunerfahrenen grundlegende Informationen über psychologische Tests zukommen** zu lassen:

- ***Testknacker***: enthalten Lösungen vieler Testitems bzw. gezielte Antwortvorschläge -> sowas ist psychologisch NICHT vertretbar!

- ***institutionalisierte und privatwirtschaftlich organisierte Seminare*** (z.B. des AMS) mit Test-Trainings für Langzeitarbeitslose ist hingegen OK!

: