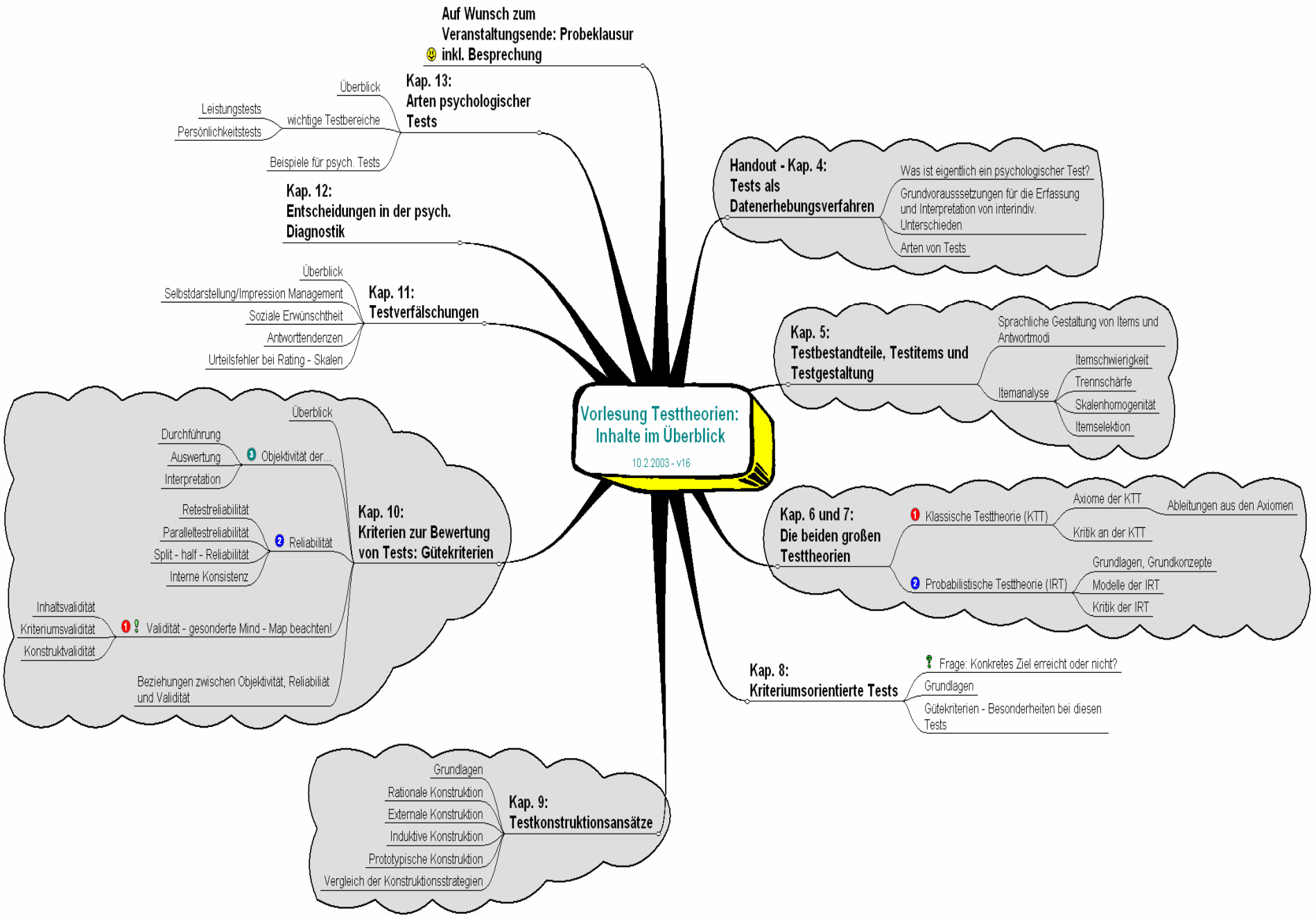
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Vorlesung Testtheorien

Dr. Tobias Constantin Haupt, MBA

Sommersemester 2007


Vorlesung Testtheorien: Inhalte im Überblick
10.2.2003 - v16
Auf Wunsch zum Veranstaltungsende: Probeklausur
inkl. Besprechung
Kap. 13: Arten psychologischer Tests
Überblick
wichtige Testbereiche
Leistungstests
Persönlichkeitstests
Beispiele für psych. Tests
Kap. 12: Entscheidungen in der psych. Diagnostik
Kap. 11: Testverfälschungen
Überblick
Selbstdarstellung/Impression Management
Soziale Erwünschtheit
Antworttendenzen
Urteilsfehler bei Rating - Skalen
Handout - Kap. 4: Tests als Datenerhebungsverfahren
Was ist eigentlich ein psychologischer Test?
Grundvoraussetzungen für die Erfassung und Interpretation von interindiv. Unterschieden
Arten von Tests
Kap. 5: Testbestandteile, Testitems und Testgestaltung
Sprachliche Gestaltung von Items und Antwortmodi
Itemanalyse
Itemschwierigkeit
Trennschärfe
Skalenhomogenität
Itemselektion
Kap. 10: Kriterien zur Bewertung von Tests: Gütekriterien
Überblick
Objektivität der...
Durchführung
Auswertung
Interpretation
Reliabilität
Retestreliabilität
Paralleltestreliabilität
Split - half - Reliabilität
Interne Konsistenz
Validität - gesonderte Mind - Map beachten!
Inhaltsvalidität
Kriteriumsvalidität
Konstruktvalidität
Beziehungen zwischen Objektivität, Reliabilität und Validität
Kap. 6 und 7: Die beiden großen Testtheorien
Klassische Testtheorie (KTT)
Axiome der KTT
Ableitungen aus den Axiomen
Kritik an der KTT
Probabilistische Testtheorie (IRT)
Grundlagen, Grundkonzepte
Modelle der IRT
Kritik der IRT
Kap. 8: Kriteriumsorientierte Tests
Frage: Konkretes Ziel erreicht oder nicht?
Grundlagen
Gütekriterien - Besonderheiten bei diesen Tests
Kap. 9: Testkonstruktionsansätze
Grundlagen
Rationale Konstruktion
Externale Konstruktion
Induktive Konstruktion
Prototypische Konstruktion
Vergleich der Konstruktionsstrategien

Leistungen der Probanden werden mit **inhaltlich** definierten **Ziele**n verglichen…
(z.B. Lehr- oder Therapiezielen).

- *Beispiele*:
  - Hat der Schüler eine best. Rechenleistung erreicht?
  - Ist ein best. Therapieziel erreicht?
  - Hat der Fahrschüler schon best. Fahrkenntnisse erworben?

…nicht mit den Normwerten einer Eichstichprobe

… daher Kontrastierung von Norm- und Kriterienorientierung

Kriterium:

- Ein **Lehrziel**, daß man erreichen kann oder auch nicht,
- Ein **Leistungskontinuum**, auf dem man unterschiedliche Positionen einnehmen kann

Normen:

- **Real**norm (Kennwerte einer Bezugsgruppe wie in der KTT) und
- **Ideal**norm
  (repräsentativer „Kanon“ von Anforderungen)
- → Kriteriumsorientierte Tests sind idealnormiert

- **Inhaltsvalidität** ist höchstes Ziel, da die Aufgaben *Stichproben des Ziel- (d.h. Kriteriums-)verhaltens* sein sollen.
- Sicherung der Inhaltsvalidität über diverse Methoden wie z. B. Expertenbefragungen und theoretische Ableitungen

Wichtige Aufgabe daher: Generierung inhaltsvalider Itemmengen

Inhaltsvalide Itemmengen liegen dann vor, wenn sie entweder

- die ***Gesamtheit*** der Kriteriumsleistungen umfassen (z.B. alle Vokabeln, die abgefragt werden sollen); dann ist die Inhaltsvalidität maximal; oder
- eine ***repräsentative Auswahl*** (am besten per Zufallsauswahl) der Kriteriumsausgaben umfaßt.

Setzung von angemessenen Normen: Idealnormen (als Zufallsstichprobe) sind

- *sachgerecht*, wenn sie nachweislich notwendig für das Erreichen nachfolgender Kriterien sind (z.B. Vorfahrtsregeln für die Führerscheinprüfung). Sie sind
- *realitätsangemessen*, wenn die Schwierigkeiten der Normen angemessen gewählt wurde (ein Fahranfänger muß seinen PKW nicht mit traumwandlerischer Sicherheit beherrschen

- Bestimmung eines *kritischen Punktwertes* (Cut-off-point)

  z.B. bei der Führerscheinprüfung (praktisch & theoretisch: wie viele und wie schwere Fehler darf man sich erlauben, um noch zu bestehen?).
- Rechtfertigung des Cut-off-points? Weshalb ist er sinnvoll gewählt?

Testgütekriterien

- Grundsätzlich dieselben wie in der KTT
- → Aber was ist hier z.B. Reliabilität?
- Und was macht man, wenn alle Probanden das Kriterium erreichen?
- Bei Nullvarianz sind die Formeln aus der KTT nicht mehr definiert.

- Alternative Gütekriterien:

Übereinstimmungs- **(Ü-) Koeffizient** von Fricke

Konfidenzintervallschätzung

- Der Standardmessfehler ist wegen der potentiellen „Nullvarianz“ nicht zu verwenden.
- Alternative z.B. Schätzung nach dem Binomialmodell.

Wichtig:

- Kriteriumsorientierte Tests sind **ideal- statt real**normorientiert.
- Auch Kriterien werden stichprobenbezogen festgelegt (wenn auch mit anderem Akzent)
- Kriteriums- und normorientierte Tests lassen sich ineinander **umwandeln**.

Zusätzliche Literatur zu kriteriumsorientiertem Testen

Bei weitergehendem Interesse…

- Klauer, K.J. (1987). *Kriteriumsorientierte Tests: Lehrbuch der Theorie und Praxis lehrzielorientierten Messens.* Göttingen: Hogrefe.
- Fisseni (1990). *Kapitel 5.*

Wie kommt man überhaupt zu einer Auswahl von Items für einen Test nach der KTT?

- Die Itemanalyse kennen Sie bereits; die Frage oben ist hier aber „strategischer“ zu verstehen, denn: wie „erzeugt“ man überhaupt einen großen Itempool, aus dem man dann die besten Items auswählt?

Testkonstruktionsansätze

Wir wollen hierzu **vier** Ansätze/ Konstruktionsstrategien betrachten:

- **Rational (deduktiv)**
- **External (kriteriumsbezogen)**
- **Induktiv (faktorenanalytisch)**
- **Prototypisch**

Definition:

- Eine Skalenkonstruktion erfolgt dann **rational**, wenn die Items aufgrund eines theoretisch fundierten und explizierten (Persönlichkeits-)Konstrukts (deduktiv) abgeleitet werden.
- *(*versus) **Intuitive** Skalenkonstruktion:
  Wenn Items aufgrund ihrer vermuteten Inhaltsvalidität zu einem theoretisch wenig explizierten Konstrukt zusammengestellt werden.

Vorgehen: Psychometrische Konstruktion:

- Vorliegen einer Theorie (z.B. Cattells Intelligenztheorie) darüber, wie sich Personen hinsichtlich bestimmter Merkmale beschreiben lassen und voneinander unterscheiden.
- Nähere Spezifizierung und Definition des interessierenden Konstrukts (z.B. Subkategorien der Intelligenz wie fluide und kristalline Intelligenz, die sich wiederum aus schlussfolgerndem Denken, Wortschatz, etc. zusammensetzen), sowie von Verhaltensindikatoren, anhand derer sich diese hypothetischen Konstrukte im Verhalten erkennen lassen (z.B. Lösen bestimmter Aufgaben).

Rationale (deduktive) Konstruktion

- Für jeden Bereich werden sodann Items in Form von Aufgaben oder Fragen nach möglichen Verhaltensweisen zu Skalen oder Subtests zusammengestellt, die als Indikatoren in Betracht kommen.
- Validierung der Skala an einem Kriterium.

Externale (kriteriumsbezogene) Konstruktion

Definition: Eine Skalenkonstruktion erfolgt dann external, wenn Items aufgrund ihrer Diskriminationsfähigkeit zwischen Mitgliedern verschiedener Gruppen (und nicht aufgrund ihrer inhaltlichen Bedeutung) zusammengestellt werden.

Vorliegen von mind. zwei Gruppen in der sozialen Realität

- zwischen denen der zu entwickelnde Test diskriminieren soll (z.B. Haupt- versus Sonderschüler oder psychisch Auffällige versus Normale).
- Wichtig: Diskriminieren, also unterscheiden, ist hier rein im Wortsinne (deskriptiv), nicht normativ im politischen gemeint!
- Den Mitgliedern der Gruppen wird eine möglichst große und inhaltlich breit gefächerte Zahl von Items vorgelegt in der Hoffnung, daß sich darunter einige befinden werden, die zwischen den Gruppen empirisch diskriminieren, also unterschiedliche Lösungswahrscheinlichkeiten zeigen.

Externale (kriteriumsbezogene) Konstruktion

- Es werden dann diejenigen (möglicherweise sehr heterogenen) Items selegiert und zu (inhaltlich nicht interpretierbaren) Skalen zusammengefaßt, die zwischen den Gruppen statistisch bedeutsam unterscheiden und bei denen diese Diskrimination in einer Kreuzvalidierung bei anderen Personen standhält.

Definition:

- Eine Skalenkonstruktion erfolgt dann **induktiv**, wenn Items (blind-analytisch) mittels einer Faktorenanalyse zu Skalen gruppiert werden, die (empirisch) hoch miteinander (und möglichst gering mit Items anderer Skalen: Einfachstruktur) korrelieren und damit gemeinsam eine Dimension konstituieren.

Vorgehen:

- Ein möglichst umfangreicher und für die Zielkonstrukte repräsentativer Itempool wird einer möglichst umfangreichen und für die Zielgruppe repräsentativen PersonenSP zur Beantwortung vorgelegt.
- Mittels einer Faktorenanalyse werden die Items zu Gruppen hoch interkorrelierender Skalen zusammengefaßt (Ziel ist eine Einfachstruktur).
- Die einzelnen Faktoren oder Skalen werden interpretiert, indem man nach einer Gemeinsamkeit aller Items einer Skala gesucht wird.

Definition:

- *Eine Skalenkonstruktion erfolgt dann **prototypisch**, wenn überwiegend solche Items zu Skalen zusammengefaßt werden, die für eine Dimension (z.B. intelligent, dominant, aggressiv) besonders (proto-)typisch oder zentral sind.*

Vorgehen:

z. B.

Act – Frequency – Approach; dt.: Handlungs-Häufigkeits-Ansatz (Buss & Craig, 1980):

- Auswählen derjenigen Eigenschaft, für die eine Skala konstruiert werden soll (z.B. Aggressivität)
- Versuchspersonen sollen an diejenigen Bekannten aus ihrem Umfeld denken, bei denen diese Eigenschaft besonders stark ausgeprägt ist.

Prototypische Konstruktion

- Versuchspersonen sollen dann konkrete Verhaltensweisen dieser Personen nennen, die indikativ für die Eigenschaft (hier: Aggressivität) sein sollen.
- Dann werden die so erhaltenen Items anderen Versuchspersonen vorgelegt, die sie nach ihrer Prototypizität hinsichtlich des Merkmals einschätzen sollen.
- D. h. also, diese Beurteiler sollen feststellen, wie prototypisch/charakteristisch die für Aggressivität gesammelten Verhaltensweiten/Acts Ihrer Meinung nach sind.

Vorteile:

- Mit prototypischen Items lassen sich kürzere Skalen konstruieren.
- Nach Prototypizitätseinschätzungen konstruierte Skalen zeigen höhere Validitäten bei Fremdeinschätzungen als Kriterium.

(möglicher) Nachteil:

- so produzierte Items sind für die Versuchperson extrem „durchschaubar“ (da sie ja gerade allgemein als prototypisch/charakteristisch für z. B. Aggressivität sind und als solche von fast jedermann erkannt werden) und damit verfälschungsgefährdet (man denke z. B. an eine Personalauswahlsituation)!

Vergleich der Konstruktionsstrategien

Kombination verschiedener Ansätze: Die verschiedenen Ansätze können hinsichtlich verschiedenster Aspekte miteinander kombiniert werden. Z.B. könnten Items rational und prototypisch erdacht, mittels der Ergebnisse einer FA bereinigt werden und dann an Extremgruppen überprüft werden.

Interne Konsistenz und Reliabilität: Da rational und induktiv entwickelte Skalen inhaltlich *homogener* sind (Items korrelieren höher miteinander) als external konstruierte Skalen, weisen sie auch eine höhere interne Konsistenz und (in der Regel) eine höhere (interne!) Reliabilität bei gleicher Testlänge auf.

Stichproben-Anfälligkeit: Insbesondere induktiv konstruierte Tests sind in ihrer Validität in hohem Maße davon abhängig, inwieweit Untersuchungs- und AnwendungsSP ähnlich zusammengesetzt sind.

Vergleich der Konstruktionsstrategien

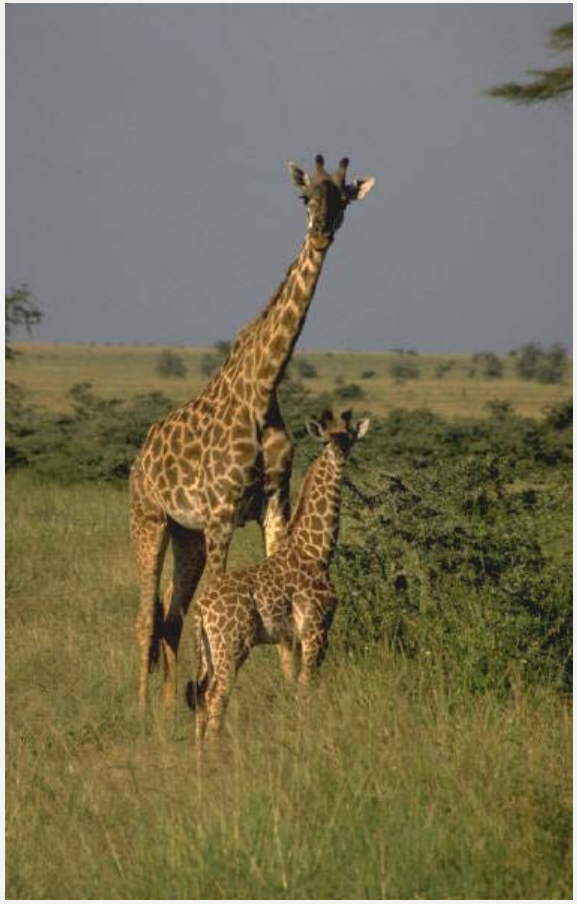
Verfälschbarkeit durch Testbeantworter: Ist insbesondere bei external konstruierten Skalen gering, da die Messintention oft nicht evident ist. Hohe Anfälligkeit für Verfälschungsversuche bei Tests, die nach dem Prototypenansatz konstruiert wurden (s. o.).

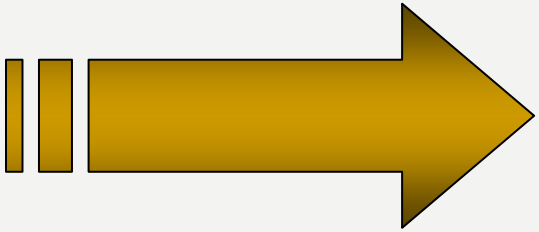
Validitäten: Es zeigt sich keine konsistente Überlegenheit von Konstruktionsstrategien gegenüber anderen.

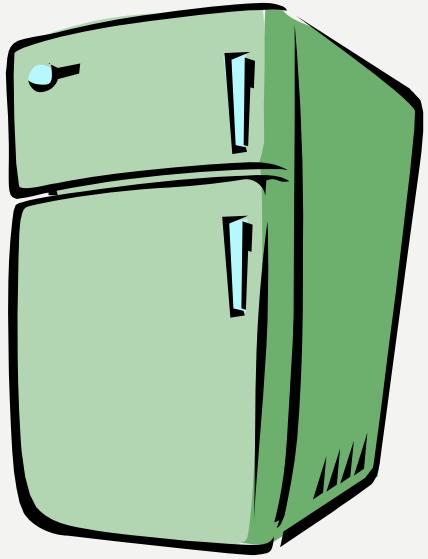
Ökonomie: Rationale Skalen sind besonders ökonomisch zu entwickeln und ihre Testergebnisse sind aufgrund der Verwendung von alltagsnahen Dimensionen leicht kommunizierbar.

The following short test consists of 4 items/ questions and tells whether you are qualified to be a "professional".

Question Number 1
How do you put a giraffe into a refrigerator?

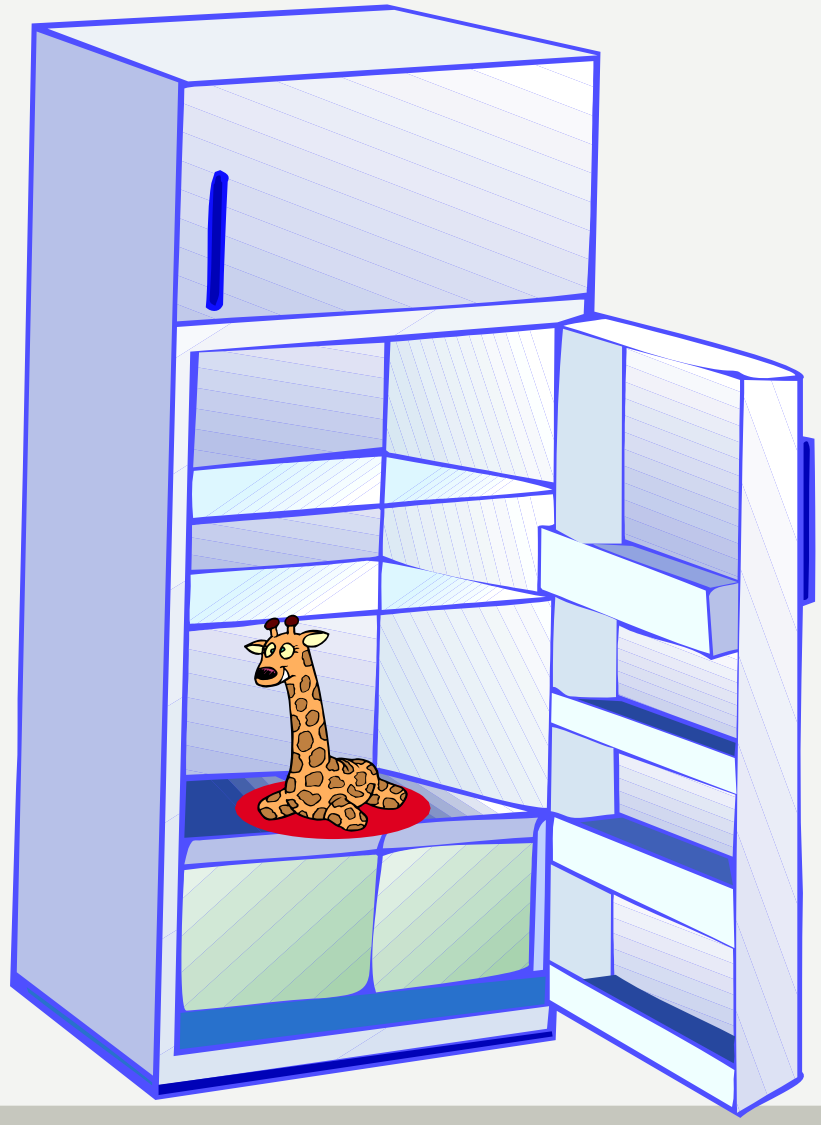






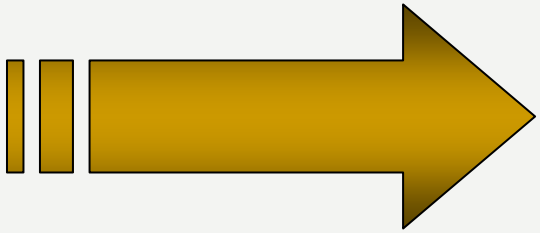
The correct answer is:

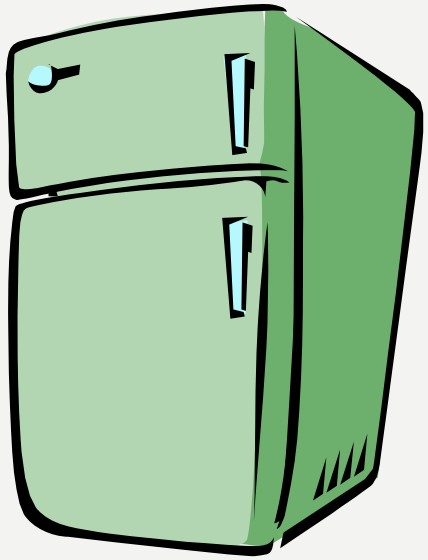
Open the refrigerator, put in the giraffe and close the door.



This question tests whether you tend to do simple things in an overly complicated way.

Question Number 2
How do you put an elephant into a refrigerator?

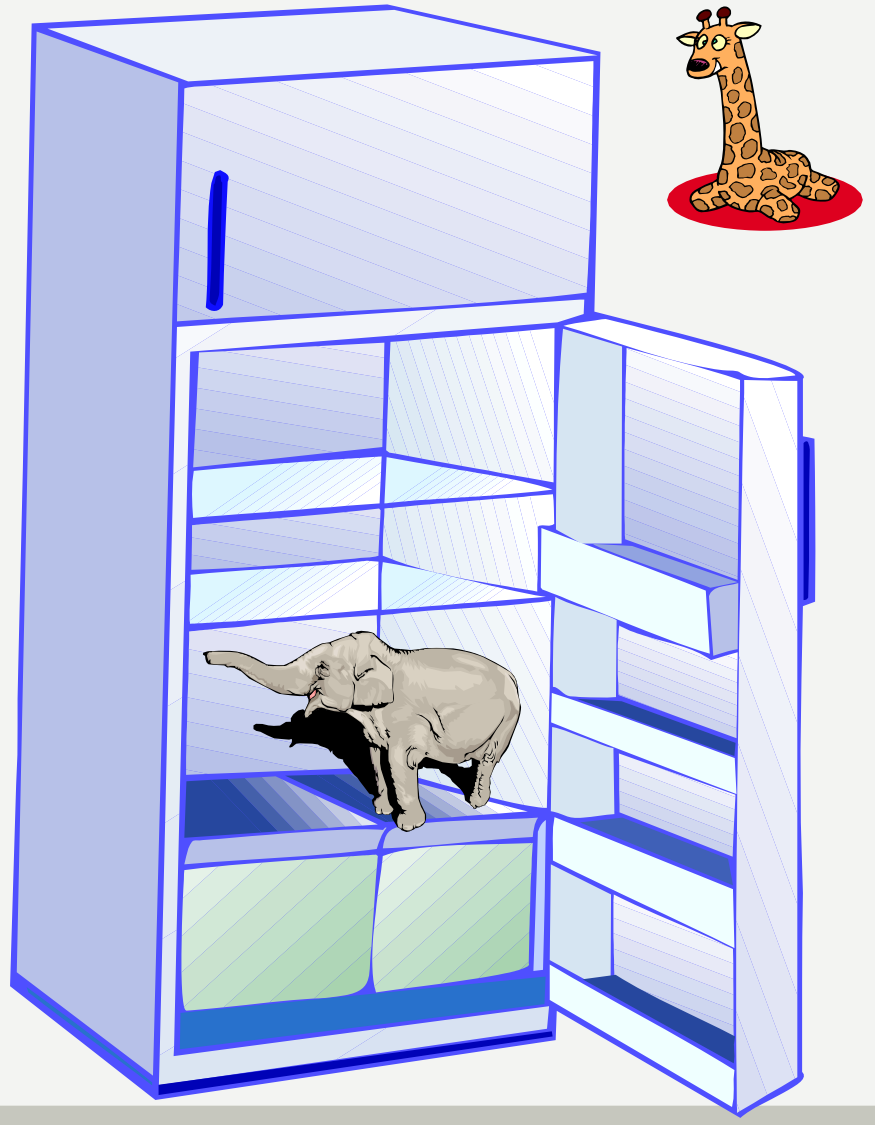




Wrong Answer: Open the refrigerator, put in the elephant and close the refrigerator.

Correct Answer:
Open the refrigerator, take out the giraffe, put in the elephant and close the door.

This tests your ability to think through the repercussions of your actions.



Question Number 3

The Lion King is hosting an animal conference. All the animals attend except one. Which animal does not attend?

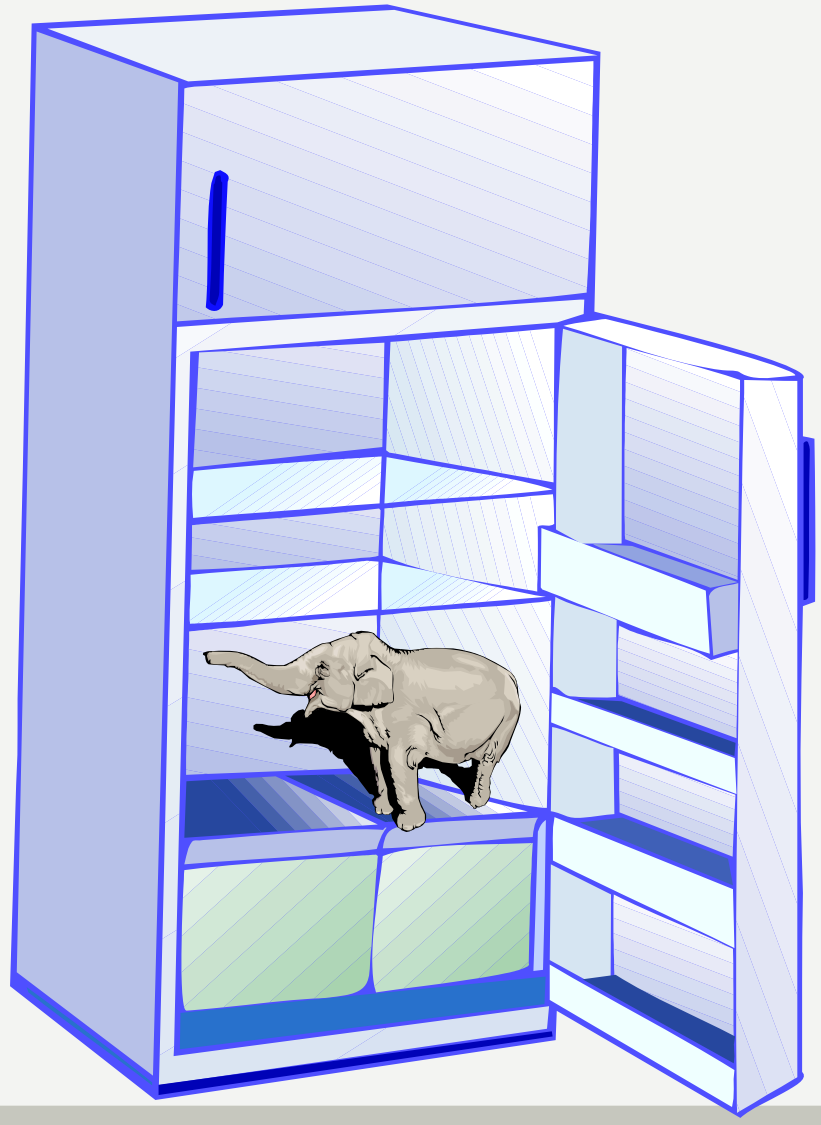
Correct Answer:

The Elephant.

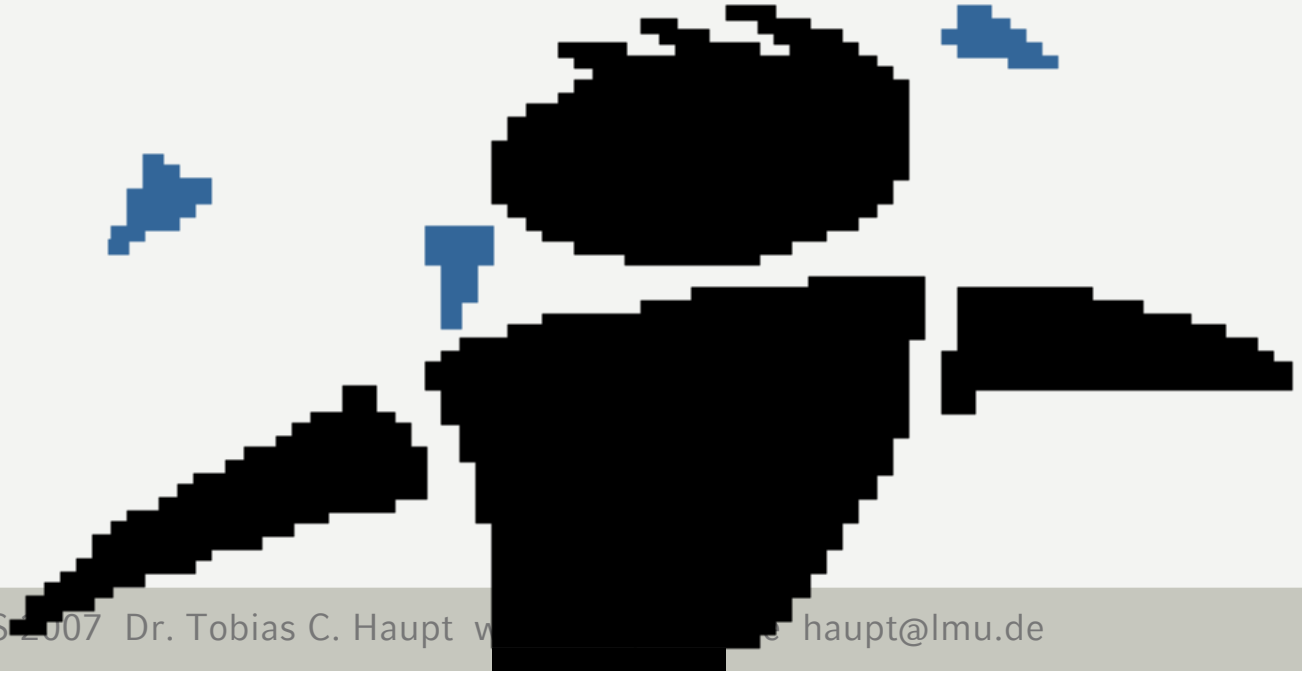
The Elephant is in the refrigerator.

Remember?

This tests your memory.



OK, even if you did not answer the first three questions correctly, you still have one more chance to show your abilities.



Question Number 4

There is a river you must cross. But crocodiles inhabit it.

How do you manage it?

Correct Answer: You swim across. Why? All the Crocodiles are attending the Animal Conference.

This tests whether you learn quickly from your mistakes.

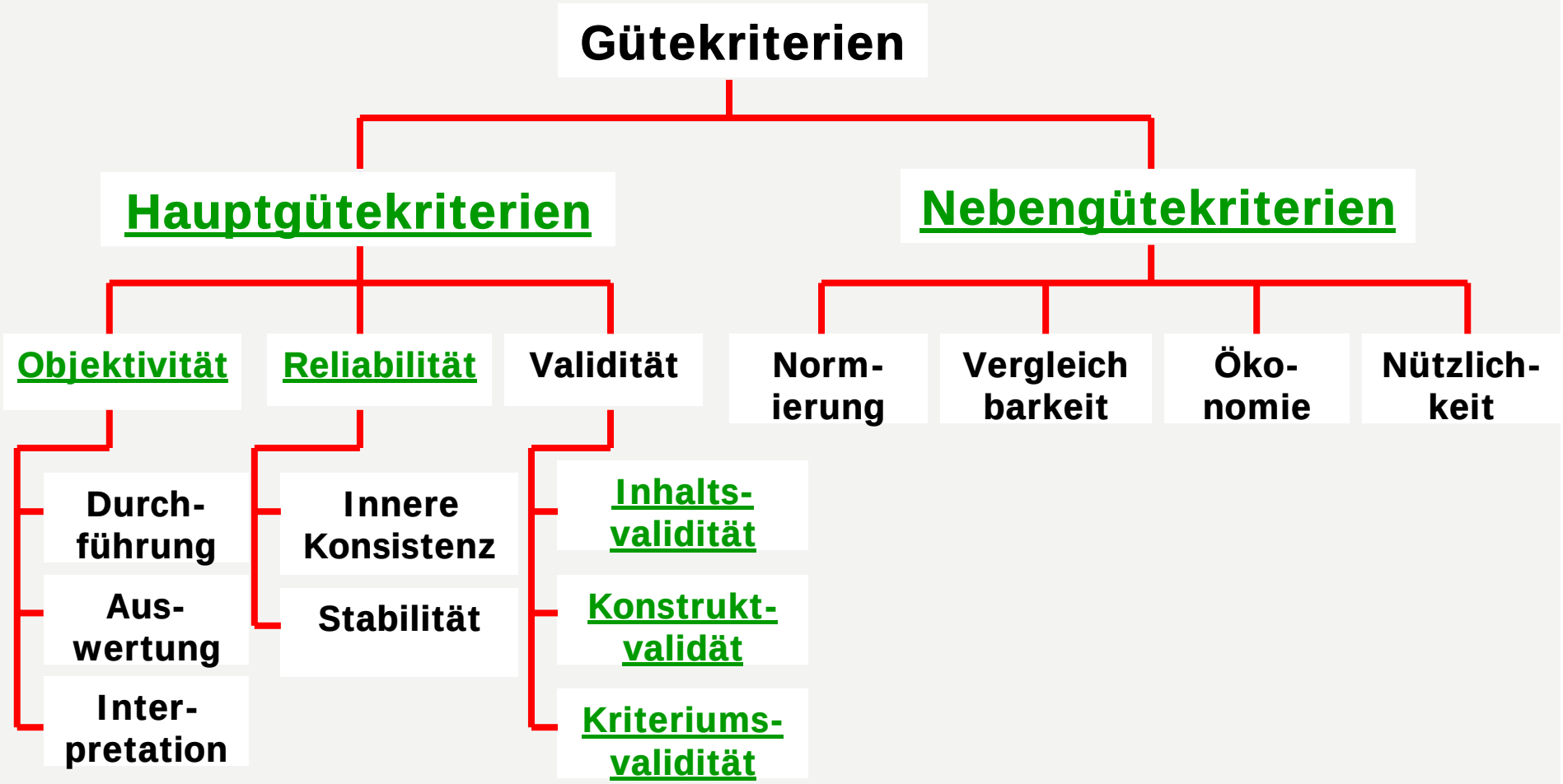
Gütekriterien

- Itemschwierigkeit, Trennschärfe und Homogenität charakterisieren einen Test von seinen kleinsten Bausteinen her (den Items).
- Den Test im Ganzen charakterisieren im Rahmen der KTT die sog. Hauptgütekriterien.
- Die grundlegende Frage dabei ist:
  Wie gut wird durch den Test das empirische Relativ (z. B. die zu erfassende Intelligenz) im numerischen Relativ (also z. B. dem IQ, der Zahl der wir der Intelligenzausprägung einer Person zuordnen) abgebildet?

Gütekriterien

- Bei der Beurteilung einer spezifischen diagnostischen Methode kommt es daher auch auf die Umstände, Bedingungen und Zielsetzungen an, was zu einer unterschiedlichen Gewichtung der Kriterien führen kann.
- Die sog. *Hauptgütekriterien* Objektivität, Reliabilität und Validität sind allerdings unter allen Umständen verbindlich.
- Von den *Nebengütekriterien* werden im weiteren Normierung (relativ testnah) und Testfairness (relativ entscheidungsnah) genauer ausgeführt werden.

Übersicht über die Haupt- und Nebengütekriterien



Die Hauptgütekriterien

- Bei der Beurteilung einer spezifischen diagnostischen Methode kommt es daher auch auf die Umstände, Bedingungen und Zielsetzungen an, was zu einer unterschiedlichen Gewichtung der Kriterien führen kann.
- Die sog. *Hauptgütekriterien* Objektivität, Reliabilität und Validität sind allerdings unter allen Umständen verbindlich.
- *Nebengütekriterien* (u. a. Normierung, Ökonomie, Nützlichkeit).

Die Objektivität

Objektivität bezeichnet das Ausmaß, in dem die Ergebnisse eines Tests (Durchführung, Auswertung, Interpretation) unabhängig vom Testleiter (Untersucher) sind.

Es lassen sich drei Objektivitätsarten unterscheiden:

- Durchführungsobjektivität
- Auswertungsobjektivität
- Interpretationsobjektivität

Definition:

Eine Testdurchführung erfolgt dann objektiv, wenn keine Testergebnisvarianz aufgrund von (möglicherweise für jede Versuchsperson unterschiedlichen) Testbedingungen und Versuchsleiter-Verhalten entsteht.

Herstellung von Durchführungsobjektivität:

durch maximale Standardisierung der Testsituation (z.B. standardisierte Instruktion, Testmaterialien, Zeitvorgaben, etc.).

Quantitative Bestimmung der Durchführungsobjektivität:

- Theoretisch müßte man eine Versuchsperson mehrmals unter denselben Bedingungen (selber Test, Versuchsleiter, etc.) testen und dann einen Mittelwert bestimmen.
- Dies ist jedoch aufgrund mangelnder Reliabilität und Testwiederholungseffekten praktisch nicht sinnvoll möglich.

Auswertungsobjektivität

Definition:

- Auswertungsobjektivität liegt vor, wenn die Vergabe von Testpunkten für best. Testantworten der Versuchspersonen unbeeinflußt von der Person des Auswerters ist.

Herstellung von Auswertungsobjektivität:

- Hohe Auswertungsobjektivität liegt in der Regel vor, wenn die Richtigkeit der Antworten mit einem Lösungsschlüssel (z.B. Schablone) oder per Computer bestimmt werden kann.
- Bei offenen Antwortformaten oder in Tests, wo mehrere Lösungen richtig sein können, bzw. auch teilweise richtig sein können oder gar bei projektiven Verfahren hängt die Auswertungsobjektivität davon ab, wie detailliert die Auswertungskategorien im Testmanual beschrieben und eingegrenzt sind.

Quantitative Bestimmung:

- indem mehrere Gutachter unabhängig voneinander das in einer Stichprobe erhobene Material auswerten und der Grad der Übereinstimmung als Korrelationskoeffizient ermittelt wird.

Interpretationsobjektivität

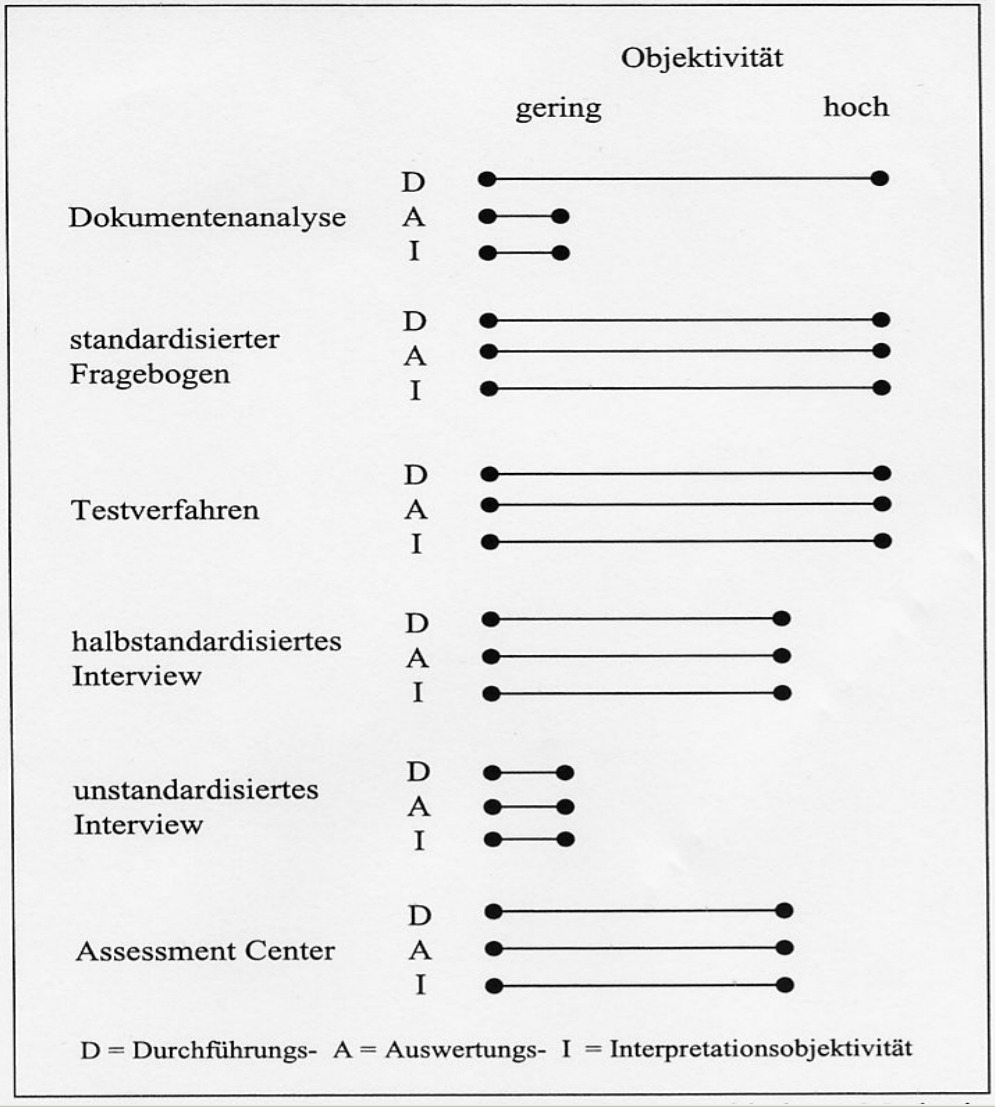
Definition:

- Interpretationsobjektivität liegt dann vor, wenn die Schlussfolgerungen (z.B. hinsichtlich der Einordnung auf einer Merkmalsdimension relativ zu anderen Versuchspersonen oder auf einer Kriteriumsdimension) unabhängig von der Person des Auswerters gezogen werden.

Herstellung von Interpretationsobjektivität:

- Hoch ist die Interpretationsobjektivität in der Regel, wenn wie in der statistischen Vorgehensweise üblich, die entsprechenden Normwerte aus einer Tabelle im Testmanual abgelesen werden können.
- In projektiven Verfahren ist die Interpretationsobjektivität dagegen meist gering, da subjektive Einschätzungen der Auswerter in die Interpretation mit eingehen.

Objektivität verschiedener Methoden



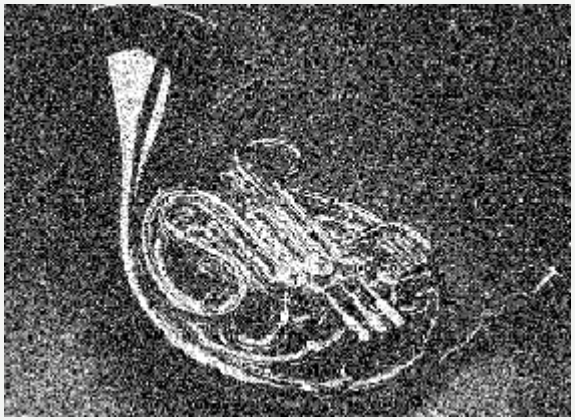
***Reliabilität** bezeichnet den…*

…Grad der Genauigkeit/die Messpräzision oder auch die Zuverlässigkeit, mit der ein Test ein bestimmtes Persönlichkeits- oder Verhaltensmerkmal misst

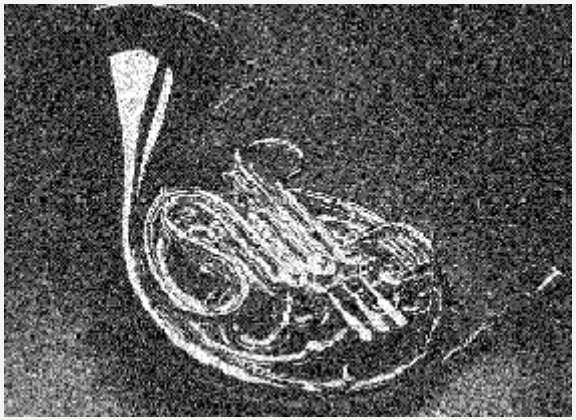
(unabhängig von der Validität, also vom Aspekt ob er das misst, was er messen soll).

Frage: Wie genau/zuverlässig/präzise erfaßt ein Test das, was er erfaßt?

Achtung: HIER IST NICHT DIE FRAGE, OB ER misst, WAS ER MESSEN SOLL!



R = .60



R = .80

R = 1.0

Haarbeispiel...

Reliabilität in der KTT:

- In der KTT wird die Annahme gemacht, daß sich der wahre Wert T zwischen zwei Messungen nicht verändert. Allgemein wird unter der Reliabilität der Anteil der wahren Varianz an der beobachteten (Gesamt-)Varianz verstanden.
- Ein guter Test sollte eine Reliabilität von über 0.8 aufweisen (= 80% der Merkmalsvarianz lassen sich auf den wahren Wert zurückführen)
- Reliabilitäten über .90 gelten als hoch.

Populationsabhängigkeit der Reliabilität

!

Beim Übergang zu homogeneren Teilpopulationen wird die Reliabilität kleiner

Die Reliabilität gibt an, wie gut ein Test in einer Bezugspopulation zu differenzieren vermag

Wie läßt sich die Reliabilität erhöhen?

- **Testlänge erhöhen**
  (Kann zu Durchführungseinschränkungen führen; Testökonomie und Zumutbarkeit reduzieren sich).
- **Homogenere Testitems verwenden**
  (damit reduzieren sich auch die Aspekte, die er erfaßt).
- **Items mittleren Schwierigkeitsgrades wählen**,
  wodurch sich auch deren Trennschärfe erhöht (wirkt allerdings einer Differenzierung in Extrembereichen entgegen).
- **Objektivität steigern**