



LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

# Vorlesung Testtheorien

Dr. Tobias Constantin Haupt, MBA

Sommersemester 2007





- **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**



- Man kann, z. B. wenn man mit der Reliabilität nicht zufrieden ist, den Test durch zusätzliche Items verlängern.
- Existiert ein Paralleltest, besteht die Möglichkeit, den Test durch parallele Items zu verdoppeln. Auch andernfalls ist immer gefordert, daß die zusätzlichen Items die gleiche Qualität wie die bisherigen haben. Sie sollten also ebenso gut in die Skala hineinpassen wie die bisherigen, v. a. eine ebenso hohe Trennschärfe besitzen.

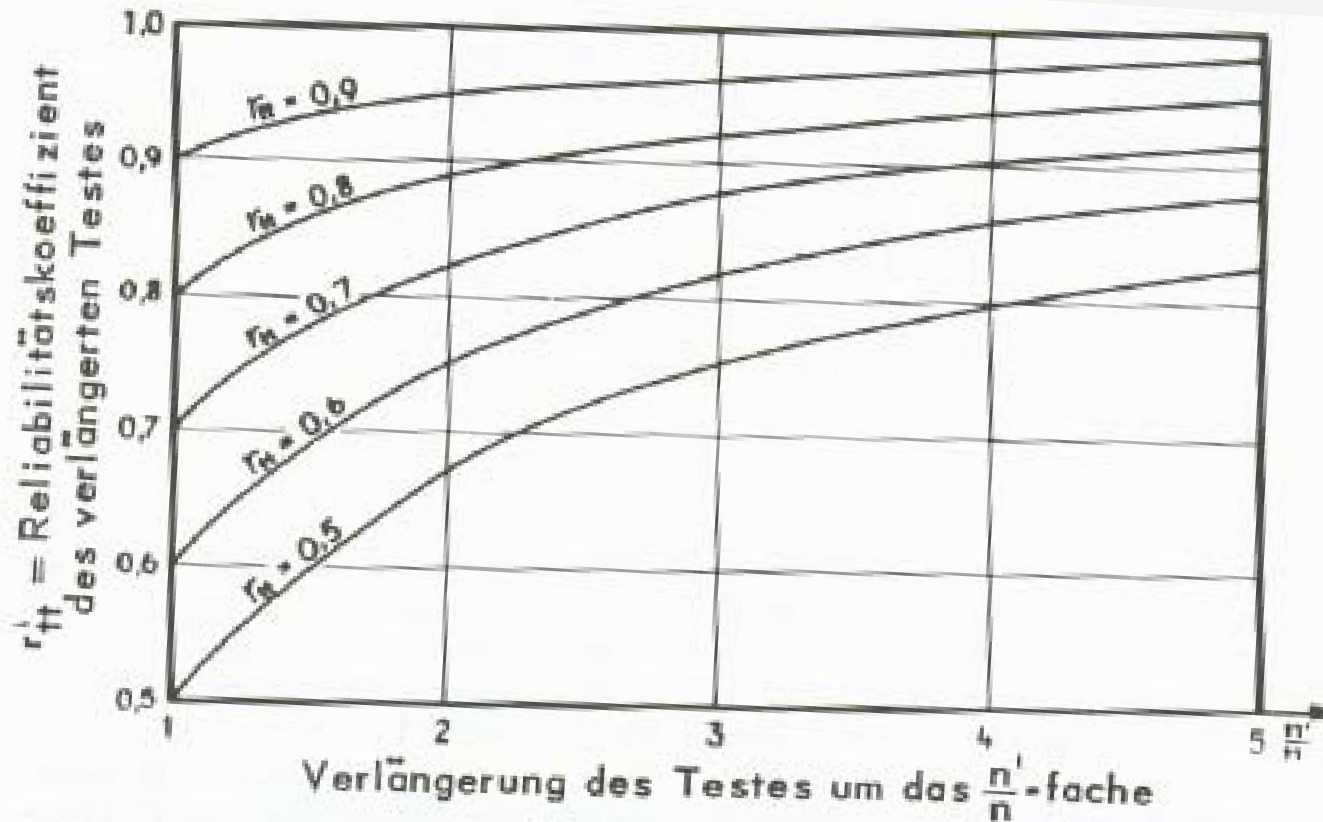


Abb. 10.3 Nomogramm zur Aufwertung eines Reliabilitätskoeffizienten bei Vervielfachung der Testlänge gemäß der verallgemeinerten SPEARMAN-BROWNSchen Formel (10.50).

## Spearman-Brown-Formel (manchmal auch "Prophecy" – Formel genannt):

$$R' = \frac{n * R}{1 + (n-1) * R}$$

$n$  = Vervielfachungsfaktor

$R$  = alte Reliabilität

$R'$  = neue Reliabilität

- Bei einer Verdoppelung setze man für  $n = 2$  ein (logo!). Wenn ich 10 Items habe und 5 sollen hinzukommen, errechnet sich  $n$  durch den Quotienten aus neuer und alter Itemanzahl (also  $15$  durch  $10 = 1.5$ ).
- Manchmal ist danach gefragt, wie die Reliabilität aussieht, wenn durch Zufall eine bestimmte Anzahl Items eliminiert würden. Angenommen man hat 10 Items und soll 3 herausnehmen, errechnet sich also  $n$  mit  $(10-3)$  geteilt durch  $10$ , also  $0,7$ .

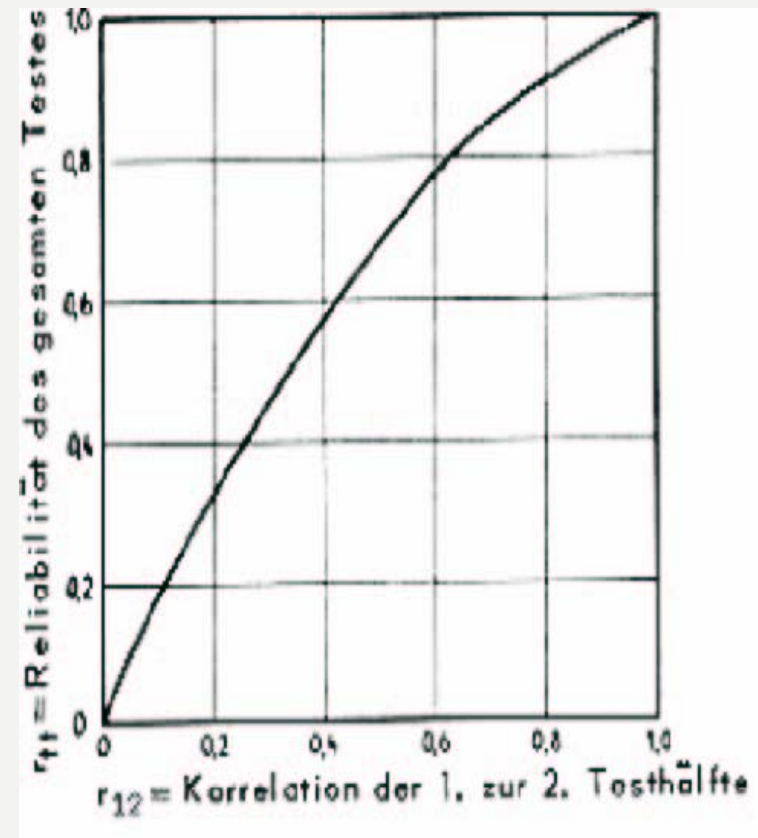
- Es gibt Aufgabentypen, bei denen es z. B. heißt, durch Testverlängerung solle die Reliabilität des 10 Items umfassenden Tests von .75 auf .90 gesteigert werden.
- Dazu muß man die obige Formel auf n auflösen:

$$n = \frac{R' * (1-R)}{R * (1-R')}$$

- Das Ergebnis gibt an, um das wie vielfache der Test verlängert werden muß, um die gewünschte Reliabilität zu erreichen. Man muß nun n noch mit der ursprünglichen Anzahl an Items multiplizieren, in unserem Beispiel  $n = 3$  multipliziert mit 10 Items. Man braucht also 30 Items. Angenommen, es käme  $n = 2,31$  heraus, ist die richtige Lösung 24 Items, da 23.1 Items nicht möglich sind - man muß dann immer aufrunden.

Wenn man über 2 Parallelförmigen eines Tests verfügt, könnte man die beiden zur Reliabilitätserhöhung auch in einem Test vereinen. Welchen Reliabilitätszuwachs man damit erzielen kann, gibt folgende Abbildung wieder.

Abb.: Nomogramm zur Aufwertung eines Reliabilitätskoeffizienten nach der Spearman – Brown – Formel.





Es gibt aber verschiedene Möglichkeiten der Reliabilitätsschätzung.

- Es gibt vier verschiedene methodische Zugänge die Reliabilität zu bestimmen:
  1. Testwiederholungsmethode
  2. Paralleltestmethode
  3. Testhalbierungsmethode
  4. Konsistenzanalyse (innere/interne Konsistenz, hier z. B. berechnet nach „Cronbachs Alpha“)



**Vorgehen:** Der gleiche Test wird denselben Versuchspersonen zu (mind.) zwei verschiedenen Zeitpunkten vorgegeben. Die Retestreliaibilität wird durch die Korrelation zwischen erster und zweiter Messung bestimmt.

**Anwendung:**

- Wenn Wiederholungseffekte (Übung, Problemeinsicht, Vertrautheit mit Testsituation, Erinnerung an Aufgaben, etc.) nicht zu erwarten sind (z.B. bei Speedtests oder Persönlichkeitstests)
- Wenn zwischen beiden Messungen ein größeres Zeitintervall liegt
- Wenn der Test beansprucht, zeitlich stabile Merkmale zu erfassen

### Nachteile:

- Überschätzung der Reliabilität, wenn Lösungen erinnert werden
- keine Anwendung bei instabilen Merkmalen (z.B. Stimmungen)
- großer untersuchungstechnischer Aufwand (da Versuchspersonen erneut kontaktiert werden müssen und mit Versuchspersonen-“Verlusten“ zu rechnen ist)

Die Reliabilität berechnet sich hier als Korrelation zweier Testdurchgänge mit den gleichen Personen.

$$R = r_{x1, x2}$$

- $x1$  = Messwert einer Person bei der ersten Messung
- $x2$  = Messwert der gleichen Person bei der zweiten Messung



**Vorgehen:** In derselben SP werden nach einer kurzen Intervallzeit zwei streng äquivalente Formen eines Tests nacheinander durchgeführt.

Parallel sind Tests dann, wenn

- Mittelwerte und Standardabweichungen identisch sind, sowie die
- Korrelationen hoch sind (beide Tests also dasselbe erfassen).

Je höher die Ergebnisse beider Tests miteinander korrelieren, um so weniger Fehlereffekte liegen vor. Die ermittelten Koeffizienten liegen meist unter denen der Retestreliaibilität.

## Anwendung:

Bei instabilen Merkmalen und wenn ohnehin zwei Testversionen entwickelt werden sollen (z.B. bei Gruppentestungen im Leistungsbereich, damit die Leute nicht voneinander in der Testsituation profitieren, also z. B. durch Spicken!).

## Vorteil:

Königsweg der Reliabilitätsbestimmung, weil ein Test so praktisch (intraindividuell) mit sich selbst verglichen werden kann, ohne daß etwa Erinnerungseffekte auftreten.

- Der „Trick“ ist also, daß man so eine Quasi – Messwiederholung ohne Transfereffekte nutzen kann.



## Probleme:

- Hoher Konstruktionsaufwand, da parallele Versionen erstellt werden müssen.
- Wenn Items einzigartig sind und nicht parallelisiert werden können.

Voraussetzung: Mittelwerte und Standardabweichungen sind identisch und Korrelation zwischen den wahren Werten der beiden Tests = 1.0.



Die Reliabilität berechnet sich hier als Korrelation der Messwerte der jeweils gleichen Person in den beiden Paralleltests.

- $R = r_{X,Y}$ 
  - $x$  = Messwert einer Person bei einem der Paralleltests
  - $y$  = Messwert der gleichen Person bei dem anderen der beiden Paralleltests

**Vorgehen:** Ein Test wird nur einmal vorgegeben und danach in zwei gleichwertige Hälften aufgeteilt

- (zum Beispiel nach der Odd-Even-Methode: die geraden Items innerhalb einer Numerierung in die eine, die ungeraden Items in die andere Hälfte.

Andere Aufteilungsmöglichkeiten sind:

- Halbierung in die erste und die zweite Hälfte (also bei 40 Items die ersten 20 und die letzten 20), die
- Auswahl per Zufall oder die
- Trennung auf der Basis von Analysedaten (Berücksichtigung von Schwierigkeit und Trennschärfe)



- ...wobei dann jede Hälfte als quasi-paralleler Teilttest zur anderen Hälfte aufgefaßt wird.
- Beide Testhälften werden dann miteinander korreliert (Testhalbierungskoeffizient, der die wahre Reliabilität unterschätzt, da die Reliabilität mit der Anzahl der Items zunimmt)...
- ...und dann nach Spearman-Brown-Formel auf die Gesamtttestlänge hochgerechnet (wichtig auch für die Klausur, dies wird oft vergessen!)



Man berechnet also zunächst  $r_{x1, x2}$

- $x1$  = Messwert einer Person bei der ersten Testhälfte
- $x2$  = Messwert der gleichen Person bei der zweiten Testhälfte

Die Spearman-Brown-Formel zur Testverdoppelung lautet:

$$R = \frac{2 * r_{x1, x2}}{1 + r_{x1, x2}}$$

Der Reliabilitätskoeffizient wird so auf die doppelte Länge der Skala aufgewertet: Korreliert werden miteinander ja nur die Hälften, doch interessant ist allein die Reliabilität der Gesamtskala.

## Anwendung:

Wenn die obigen Reliabilitätsbestimmungen nicht durchgeführt werden können.

## Vorteile:

- geringer untersuchungstechnischer Aufwand
- Motivations-, Stimmungs-, Aufmerksamkeitschwankungen können als Varianzquelle praktisch ausgeschlossen werden, so daß die instrumentelle Messgenauigkeit hier am besten bestimmt werden kann.



- Erweiterung, bzw. Verallgemeinerung der Testhalbierungsmethode, wobei **jedes** Item als ein Paralleltest behandelt wird (setzt Homogenität, bzw. Äquivalenz der Items voraus).
- Die Analyse erfolgt analog zur Bestimmung der Split-half-Reliabilität, nur daß hier der Test nicht in zwei Teile, sondern in so viele Teile, wie er Items hat, zerlegt wird. Die Analyse liefert dann einen mittleren Split-Half-Koeffizienten.

Berechnung: kann entweder

- bei dichotomen Antwortformaten durch Korrelationen zwischen allen Items bestimmt werden (mittels einer Kuder-Richardson-Formel), welche dann die „wahre“ Varianz widerspiegeln



...oder:

bei mehrstufigen Antwortformaten (häufiger verwendet) durch Cronbachs Alpha (Koeffizient entspricht der mittleren Testhalbierungsreliabilität eines Tests für alle möglichen Testhalbierungen)

Bei heterogenen Tests unterschätzt  $\alpha$  jedoch die Reliabilität und wird daher auch als Homogenitätsindex bezeichnet.

## Vorteil:

Führt zu stabileren Schätzungen als die Testhalbierungsmethode, bei welcher die Koeffizienten von der Halbierungsmethode (Sie wissen ja: z. B. odd-even; Zufall; ....) abhängen und ist daher zu bevorzugen.

Dies ist quasi eine Verallgemeinerung der Halbierungsmethode. Rein theoretisch wird hier die Testskala nicht nur in zwei Hälften zerlegt, sondern in so viele wie Items vorhanden sind (wie gesagt).



## Berechnung von Cronbachs $\alpha$ :

$$\alpha = \frac{n}{n-1} * \left( 1 - \frac{\sum_{i=1}^n s_i^2}{s_{\text{sum}}^2} \right)$$

- $\alpha$  = Wert für die innere Konsistenz der Skala
- $n$  = Anzahl der Items
- $s_i^2$  = Varianz der Werte des „i-ten“ Items.
- $s_{\text{sum}}^2$  = Varianz der Summenwerte/Testergebnisse der Personen

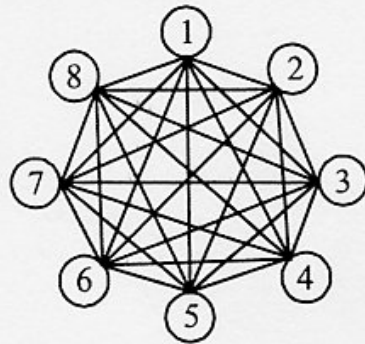


Übersicht: Arten von Tests und Methoden der Reliabilitätsschätzung

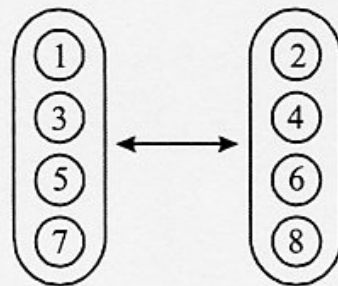
	Speed-Tests	Power-Tests	homogene Tests	heterogene Tests
Retest	ja	(nein)	ja	ja
Paralleltest	ja	ja	ja	(ja)
Interne Konsistenz	nein	ja	ja	nein



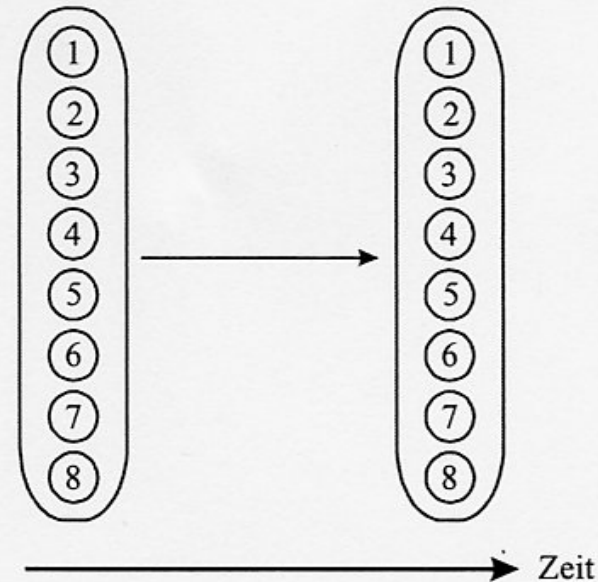
a) Innere Konsistenz



b) Split-Half-Reliabilität



c) Retest-Reliabilität







**Reliabilitätsanalyse: Statistik** ✖

Deskriptive Statistiken für

- Item
- Skala
- Skala wenn Item gelöscht

Zwischen Items

- Korrelationen
- Kovarianzen

Auswertung

- Mittelwert
- Varianzen
- Kovarianzen
- Korrelationen

ANOVA-Tabelle

- Keine
- F-Test
- Friedman Chi-Qadrat
- Cochran Chi-Qadrat

Hotellings T-Qadrat       Tukeys Additivitätstest

Korrelationskoeffizient in Klassen

Modell:  Typ:

Konfidenzintervall:  %      Testwert:



RELIABILITY ANALYSIS - SCALE (ALPHA)

		Mean	Std Dev	Cases
1.	N2	2,2772	,9604	101,0
2.	N7	2,8119	,9134	101,0
3.	N12	2,6535	,9740	101,0
4.	N17	3,1287	,8326	101,0
5.	N22	1,8317	1,0590	101,0
6.	N27	2,0000	,9798	101,0
7.	N32	2,1089	,8706	101,0
8.	N37	2,6634	,8158	101,0
9.	N42	2,3465	1,0811	101,0
10.	N47	1,6832	1,0481	101,0
11.	N52	2,5842	,8279	101,0
12.	N57	1,9307	,9618	101,0

N of Cases = 101,0

Inter-item Correlations	Mean	Minimum	Maximum	Range	Max/Min	Variance
	,1880	-,2904	,4397	,7302	-1,5140	,0182

- Die mittlere Interitemkorrelation stellt einen Homogenitätsindex dar. Es handelt sich um die mittlere Korrelation zwischen den Items, diese ist nicht (!) Fisher-Z-transformiert



## Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item- Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
N2	25,7426	25,6731	,5738	,3671	,6776
N7	25,2079	29,3263	,2041	,1907	,7269
N12	25,3663	27,2945	,3868	,2353	,7037
N17	24,8911	27,7180	,4298	,2708	,6996
N22	26,1881	28,0943	,2640	,1887	,7218
N27	26,0198	26,1596	,5049	,3280	,6870
N32	25,9109	28,3220	,3345	,1429	,7107
N37	25,3564	26,7117	,5695	,4306	,6834
N42	25,6733	27,0622	,3511	,3779	,7093
N47	26,3366	30,2455	,0725	,2579	,7478
N52	25,4356	27,8283	,4197	,2106	,7009
N57	26,0891	28,1420	,3050	,1574	,7147

- Die „Corrected Item-Total Correlation“ ist die Part-Whole korrigierte Trennschärfe ( $r_{it}$ )



Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item- Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
N2	25,7426	25,6731	,5738	,3671	,6776
N7	25,2079	29,3263	,2041	,1907	,7269
N12	25,3663	27,2945	,3868	,2353	,7037
N17	24,8911	27,7180	,4298	,2708	,6996
N22	26,1881	28,0943	,2640	,1887	,7218
N27	26,0198	26,1596	,5049	,3280	,6870
N32	25,9109	28,3220	,3345	,1429	,7107
N37	25,3564	26,7117	,5695	,4306	,6834
N42	25,6733	27,0622	,3511	,3779	,7093
N47	26,3366	30,2455	,0725	,2579	,7478
N52	25,4356	27,8283	,4197	,2106	,7009
N57	26,0891	28,1420	,3050	,1574	,7147

- Die „Squared Multiple Correlation“ ist die quadrierte multiple Korrelation der restlichen Items mit dem vorliegenden Item
- Je höher die multiple Korrelation ist, desto repräsentativer ist dieses Item für alle anderen Items, da der Anteil der gemeinsamen Varianz dann höher ist



## Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item- Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
N2	25,7426	25,6731	,5738	,3671	,6776
N7	25,2079	29,3263	,2041	,1907	,7269
N12	25,3663	27,2945	,3868	,2353	,7037
N17	24,8911	27,7180	,4298	,2708	,6996
N22	26,1881	28,0943	,2640	,1887	,7218
N27	26,0198	26,1596	,5049	,3280	,6870
N32	25,9109	28,3220	,3345	,1429	,7107
N37	25,3564	26,7117	,5695	,4306	,6834
N42	25,6733	27,0622	,3511	,3779	,7093
N47	26,3366	30,2455	,0725	,2579	,7478
N52	25,4356	27,8283	,4197	,2106	,7009
N57	26,0891	28,1420	,3050	,1574	,7147

- „Alpha if Item deleted“ kennzeichnet die Reliabilität der verbleibenden Items, nachdem das vorliegende Item ausgesondert wurde

R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   ( A L P H A )

Reliability Coefficients      12 items

Alpha = ,7254      Standardized item alpha = ,7353

- Unter „Alpha“ ist die Gesamtreliabilität unter Berücksichtigung aller Items angegeben
- Das standardisierte alpha basiert anders als alpha auf Korrelationen (vor der Berechnung werden alle Items z-standardisiert) und nicht auf Varianzen
- Es führt bei ähnlichen Itemvarianzen zu geringfügig anderen Ergebnissen als Cronbach-alpha
- Sind die Itemvarianzen stark unterschiedlich, sollte das standardisierte alpha verwendet werden





- Die wichtigsten Ergebnisse der Reliabilitätsanalyse lassen sich folgendermaßen beschreiben:
- Das Alpha ist, wenn wir nach der Tabelle gehen, gering „Alpha“ = .7254, ebenso wie die mittlere Interitemkorrelation „Inter-item correlations“ (MIC) = .1880
- Besonders fällt die geringe Trennschärfe des Items N47 mit  $\tilde{n}$  „Corrected Item-Total Correlation“ = .0725 ins Auge
- Die Eliminierung dieses Items hätte eine Erhöhung der Reliabilität auf „Alpha if item deleted“ = .7478 zur Folge
- Neben möglichen inhaltlichen Gründen für die geringe Trennschärfe sind zwei weitere Gründe denkbar. Zum einen handelt es sich bei dem Item N47 um das Item mit der extremsten Schwierigkeit ( $M = 1.6832$ ) im Vergleich zu den anderen Itemmittelwerten, zum anderen ist es im Gegensatz zu allen anderen Items deutlich linkssteil verteilt

- Im Folgenden werden die Ergebnisse der erneuten Reliabilitätsanalyse ohne das Item N47 und ohne Mittelwerte und Standardabweichungen beschrieben
- Es empfiehlt sich, bei der so genannten alpha-Maximierung sukzessiv vorzugehen
- Das heißt, es wird zunächst nur ein Item aus der Analyse ausgeschlossen
- Dadurch ändern sich die Summenwerte der Skala und auch alle Trennschärfen
- So kann es möglich sein, dass nach Aussonderung des Items N47 das Item N7 eine höhere Trennschärfe erhält und das alpha nach Elimination des Items nicht mehr ansteigt



```

Inter-item
Correlations          Mean      Minimum      Maximum      Range      Max/Min      Variance
                    ,2171      -,0951       ,4397       ,5348      -4,6241      ,0126
    
```

### Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item- Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
N2	24,0594	23,8964	,5779	,3650	,7038
N7	23,5248	27,2519	,2264	,1837	,7503
N12	23,6832	25,3186	,4059	,2353	,7278
N17	23,2079	25,7463	,4504	,2697	,7232
N22	24,5050	26,6525	,2261	,1623	,7542
N27	24,3366	24,5255	,4905	,3074	,7159
N32	24,2277	26,6776	,3124	,1303	,7394
N37	23,6733	24,7822	,5907	,4294	,7068
N42	23,9901	24,4699	,4307	,2639	,7245
N52	23,7525	26,2881	,3854	,1833	,7308
N57	24,4059	26,2836	,3079	,1556	,7409

```
Reliability Coefficients      11 items
```

```
Alpha = ,7478      Standardized item alpha = ,7531
```



## Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item- Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
N2	24,0594	23,8964	,5779	,3650	,7038
N7	23,5248	27,2519	,2264	,1837	,7503
N12	23,6832	25,3186	,4059	,2353	,7278
N17	23,2079	25,7463	,4504	,2697	,7232
N22	24,5050	26,6525	,2261	,1623	,7542
N27	24,3366	24,5255	,4905	,3074	,7159
N32	24,2277	26,6776	,3124	,1303	,7394
N37	23,6733	24,7822	,5907	,4294	,7068
N42	23,9901	24,4699	,4307	,2639	,7245
N52	23,7525	26,2881	,3854	,1833	,7308
N57	24,4059	26,2836	,3079	,1556	,7409

- Anhand dieser Itemanalyse könnten aufgrund ihrer „relativ“ niedrigen Trennschärfe zwei weitere Items N7 und N22 ausgesondert werden
- Doch hier ist Vorsicht geboten, denn ein Test kann auch „zu Tode“ homogenisiert oder optimiert werden



## R E L I A B I L I T Y   A N A L Y S I S   -   S C A L E   ( A L P H A )

		Mean	Std Dev	Cases
1.	N2	2,2772	,9604	101,0
2.	N7	2,8119	,9134	101,0
3.	N12	2,6535	,9740	101,0
4.	N17	3,1287	,8326	101,0
5.	N22	1,8317	1,0590	101,0
6.	N27	2,0000	,9798	101,0
7.	N32	2,1089	,8706	101,0
8.	N37	2,6634	,8158	101,0
9.	N42	2,3465	1,0811	101,0
10.	N47	1,6832	1,0481	101,0
11.	N52	2,5842	,8279	101,0
12.	N57	1,9307	,9618	101,0

N of Cases = 101,0



## RELIABILITY ANALYSIS - SCALE (ALPHA)

		Mean	Std Dev	Cases
1.	N2	2,2772	,9604	101,0
2.	N7	2,8119	,9134	101,0
3.	N12	2,6535	,9740	101,0
4.	N17	3,1287	,8326	101,0
5.	N22	1,8317	1,0590	101,0
6.	N27	2,0000	,9798	101,0
7.	N32	2,1089	,8706	101,0
8.	N37	2,6634	,8158	101,0
9.	N42	2,3465	1,0811	101,0
10.	N47	1,6832	1,0481	101,0
11.	N52	2,5842	,8279	101,0
12.	N57	1,9307	,9618	101,0
N of Cases =		101,0		

Die Itemschwierigkeit variiert von 1.68 bis 3.13

- Es werden etwa 29 Prozent (Range:  $3.13 - 1.68 = 1.44$ ) der theoretisch möglichen Breite der 5-stufigen Antwortskala ausgeschöpft, dabei zu einem größeren Anteil die rechte Skalenhälfte (Zustimmung)



## RELIABILITY ANALYSIS - SCALE (ALPHA)

		Mean	Std Dev	Cases
1.	N2	2,2772	,9604	101,0
2.	N7	2,8119	,9134	101,0
3.	N12	2,6535	,9740	101,0
4.	N17	3,1287	,8326	101,0
5.	N22	1,8317	1,0590	101,0
6.	N27	2,0000	,9798	101,0
7.	N32	2,1089	,8706	101,0
8.	N37	2,6634	,8158	101,0
9.	N42	2,3465	1,0811	101,0
10.	N47	1,6832	1,0481	101,0
11.	N52	2,5842	,8279	101,0
12.	N57	1,9307	,9618	101,0
N of Cases =		101,0		

Die Itemstreuungen bewegen sich in einem Bereich von .82 bis 1.08



## Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item- Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
N2	25,7426	25,6731	,5738	,3671	,6776
N7	25,2079	29,3263	,2041	,1907	,7269
N12	25,3663	27,2945	,3868	,2353	,7037
N17	24,8911	27,7180	,4298	,2708	,6996
N22	26,1881	28,0943	,2640	,1887	,7218
N27	26,0198	26,1596	,5049	,3280	,6870
N32	25,9109	28,3220	,3345	,1429	,7107
N37	25,3564	26,7117	,5695	,4306	,6834
N42	25,6733	27,0622	,3511	,3779	,7093
N47	26,3366	30,2455	,0725	,2579	,7478
N52	25,4356	27,8283	,4197	,2106	,7009
N57	26,0891	28,1420	,3050	,1574	,7147

Reliability Coefficients 12 items

Alpha = ,7254 Standardized item alpha = ,7353

Item N47 weist eine sehr geringe Trennschärfe („Corrected item total correlation“) auf

- Diese geringe Trennschärfe wirkt sich auch auf die Messgenauigkeit der Skala („Alpha“) negativ aus: Das Cronbach-alpha steigt an, wenn man das Item nicht in die Itemanalyse aufnehmen würde („alpha if item deleted“)





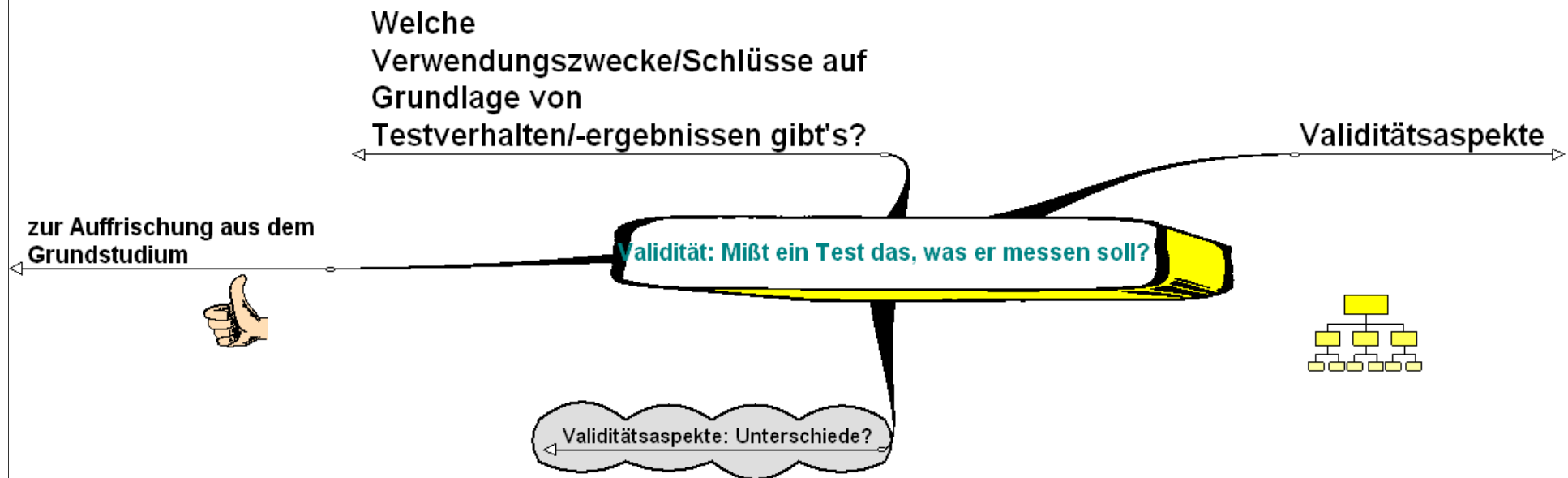
## Definition:

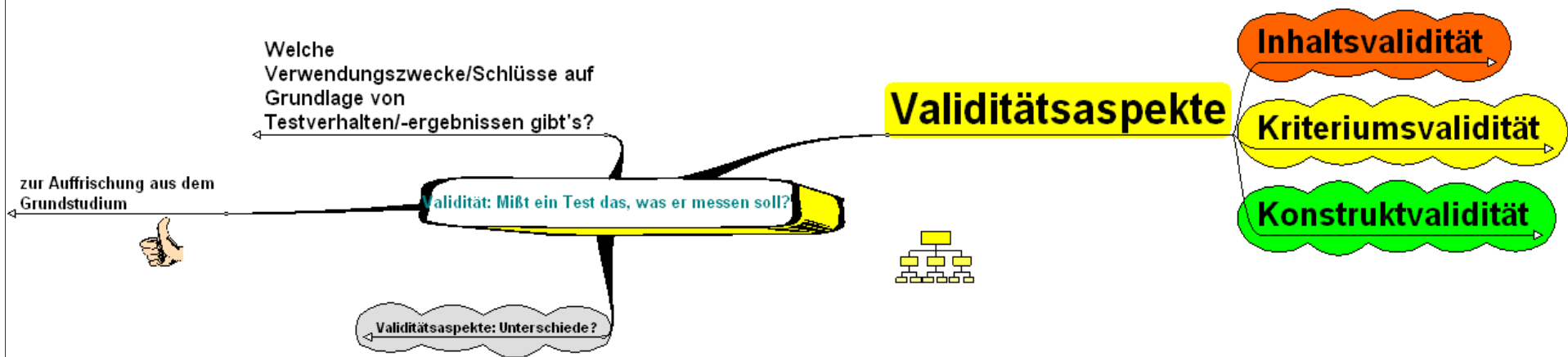
- Unter Validität versteht man den Grad der Genauigkeit, in dem ein Test (oder ein Verfahren) wirklich das misst (z. B. Persönlichkeitsmerkmal, Verhaltensweise), was er zu messen vorgibt oder messen soll.
- Ist die Validität hoch, kann von Testergebnissen (numerisches Relativ) auf Verhalten außerhalb der Testsituation (empirisches Relativ) generalisiert werden.

Was hat Validität mit Religion zu tun ???

**wichtigstes Gütekriterium:** Denn ein Test kann trotz hoher Objektivität und hoher Reliabilität (die nur günstige Voraussetzungen zur Erreichung hoher Validität schaffen) unbrauchbar sein, wenn er nicht das misst, was er messen soll

- (z.B. können Reaktionszeitmessungen am Computer in einer verkehrspsychologischen Testbatterie hoch reliabel sein, aber nichts über Reaktionen im Straßenverkehr aussagen).







- Inwiefern repräsentiert das Messverfahren das zu messende unmittelbar?
- Kein mathemat./numerischer Kennwert!
- Schluss auf Verhalten außerhalb des Tests, das dem Verhalten im Test ähnlich ist (ohne sich auf ein latentes Merkmal/Konstrukt zu beziehen)



## Definition: Inhaltsvalidität liegt vor, wenn

- der Inhalt der Testitems das zu messende Konstrukt in seinen wichtigsten Aspekten erschöpfend erfaßt oder wenn
- die Aufgaben eines Tests (repräsentative) Stichproben aus dem zu erfassenden Zielmerkmal selbst darstellen  
(Repräsentationsschluss von Testsituation auf Realsituation). Die Inhaltsvalidität ist maximal, wenn Testsituation und Kriterium vollständig (und augenfällig) übereinstimmen.
- Beispiel: „Englischer Wortschatz“

Numerische(r) Kennwert(e), so viele wie es sinnvolle Kriterien gibt!

- mögliches Problem: diese Validitätsart kann nicht unabhängig von den Reliabilitäten von Test und Kriterium ermittelt werden; evtl. Minderungskorrektur einsetzen bei Interesse an theoret./"wahren"/"was wäre bei perfekter Reliabilität" - Zusammenhängen!
  - für praktische (Vorhersage-) Zwecke jedoch unkorrigierte Werte nehmen!