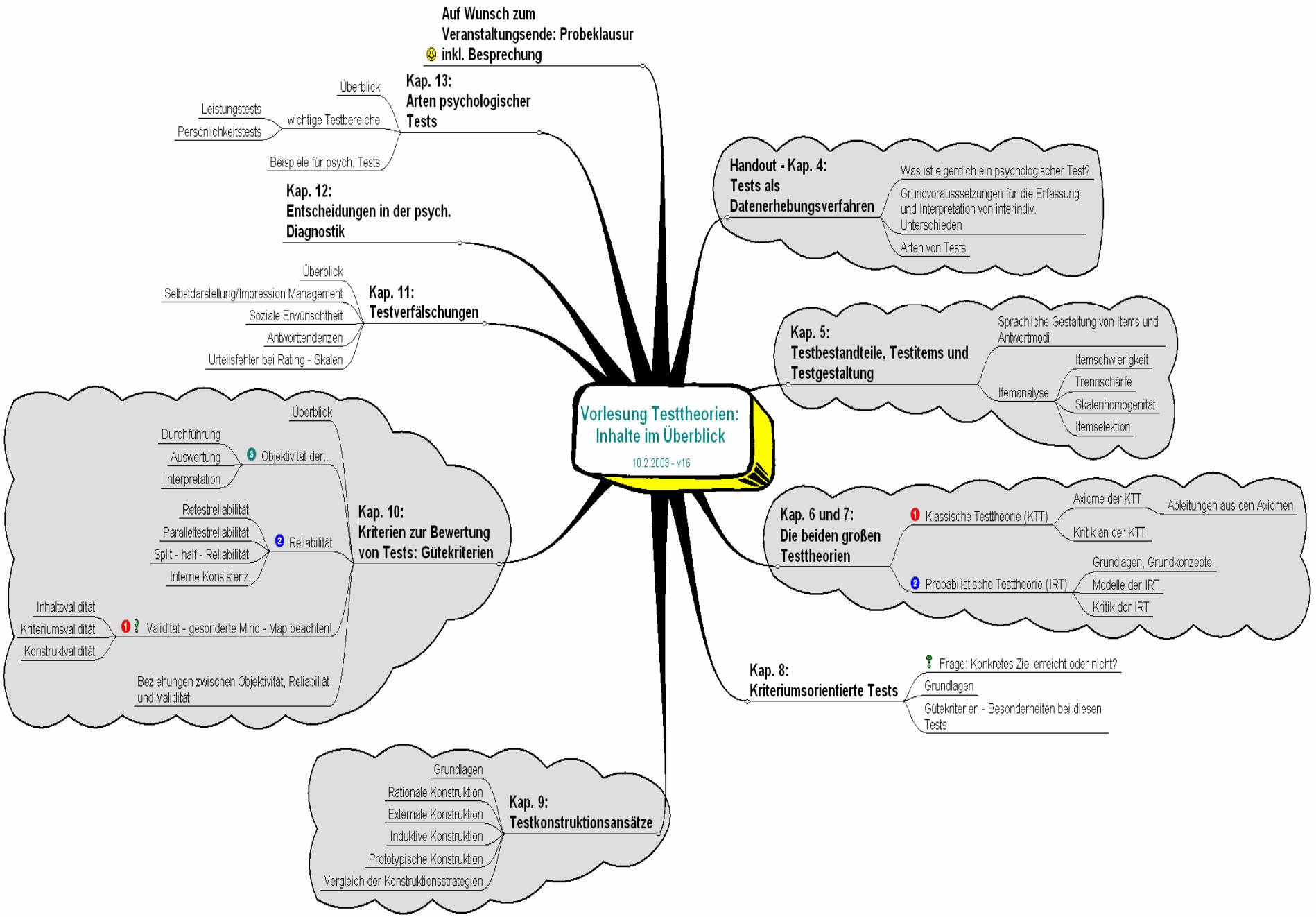
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

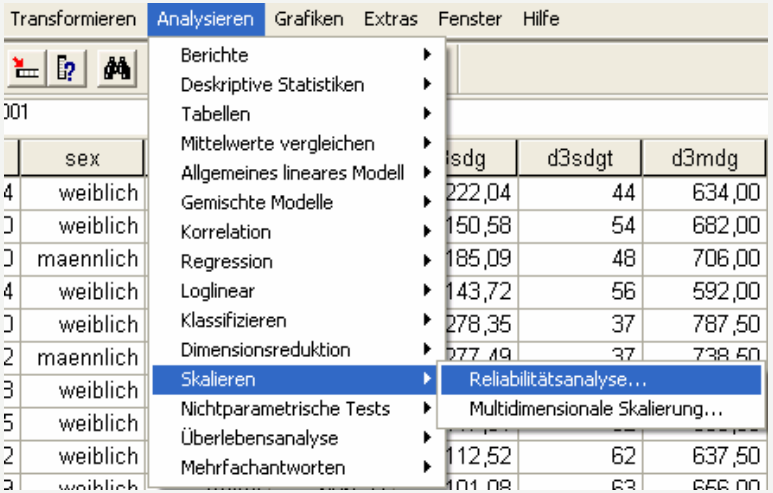
Vorlesung Testtheorien

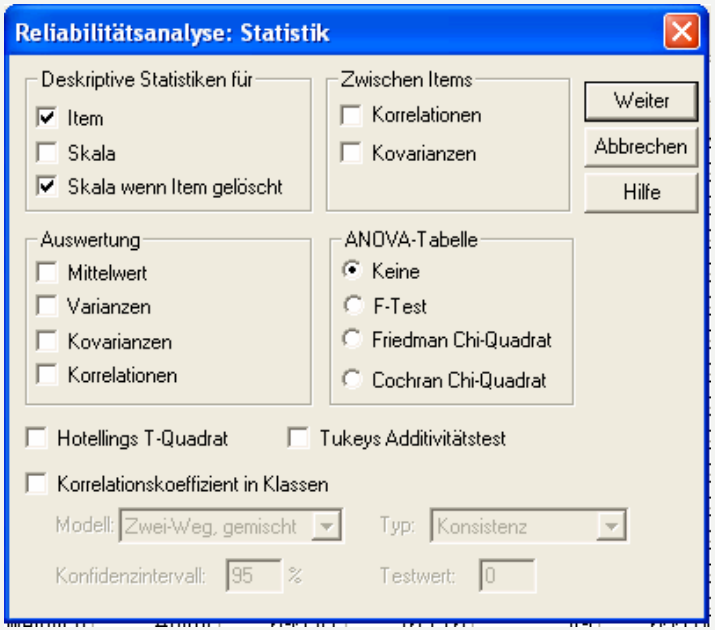
Dr. Tobias Constantin Haupt, MBA

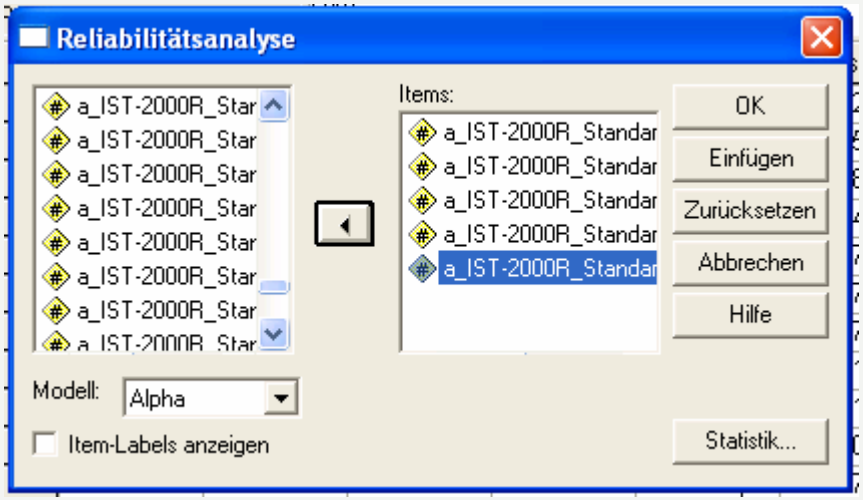
Sommersemester 2007


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

Berechnung von Trennschärfen mit SPSS







Das Kriterium *Schwierigkeit*

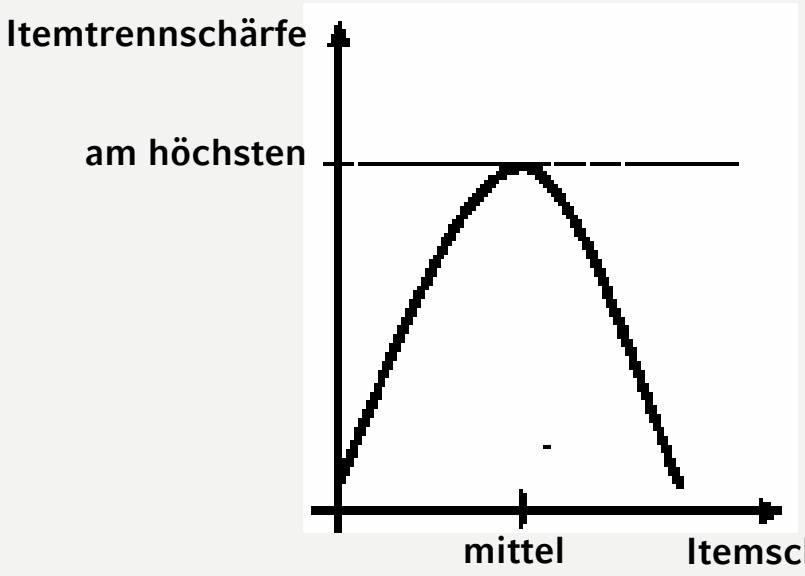
Bei einem Niveautest sollten die Schwierigkeiten entweder bei 50% (enger Geltungsbereich) liegen oder in einem großen Bereich um 50% streuen (weiter Geltungsbereich). Dabei sollten sie über den Bereich des gesamten Merkmals streuen (in einem Bereich von > 20% und < 80% der Skala).

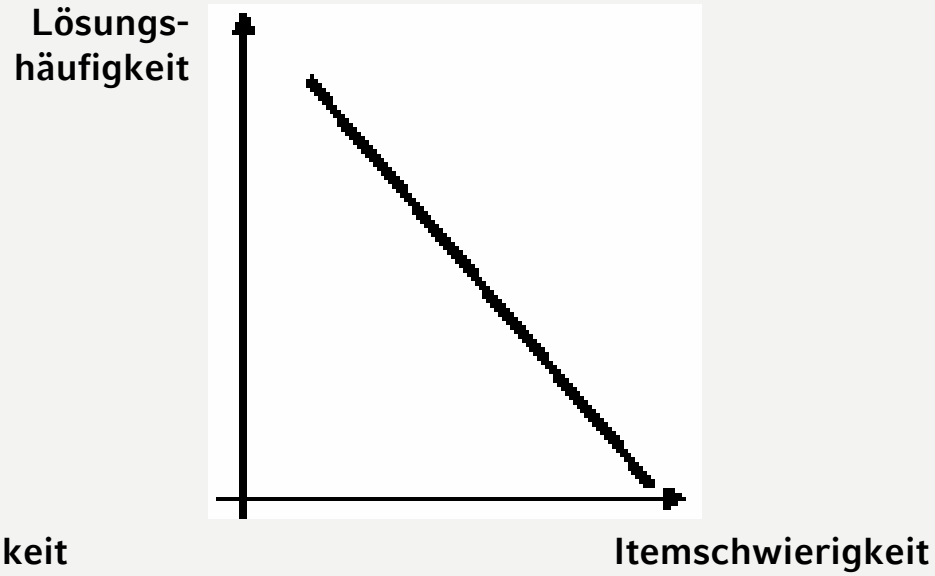
Man sollte so viele Aufgaben mit geringer Schwierigkeit erhalten, dass jeder Proband noch Punkte bekommen kann, und so viele mit hoher Schwierigkeit, dass nur wenige Probanden alle Aufgaben lösen.

Schwierigkeitsindizes sollten sich an der Stelle der Schwierigkeitsskala häufen, an der von dem Test eine besonders gute Differenzierung verlangt wird. Wenn eine gleichmäßige Differenzierung verlangt wird, sollten sie sich etwa normal mit dem Gipfel in der Mitte der Skala verteilen.

Empirischer Zusammenhang Itemschwierigkeit – Trennschärfe

Zu leichte Items werden von fast allen Probanden gelöst und zu schwere Items von keinem Probanden. Das Item trennt also am besten („am schärfsten“) bei mittlerer Itemschwierigkeit.





Itemanalyse – Aufgaben-Interkorrelation

Die Berechnung der Aufgaben-Interkorrelation ist – anders als die Berechnung der Itemschwierigkeit und Trennschärfe - zur Testkonstruktion nicht unbedingt notwendig. Sie liefert aber einen Überblick über alle Zusammenhänge zwischen jeweils zwei Items. An ihr kann man entscheiden, inwieweit zwei Items dasselbe Konstrukt messen.

Sie wird mit dem entsprechenden Korrelationskoeffizienten berechnet und als Matrix dargestellt.

Itemselektion: Weitere Kriterien

- Häufen sich Items in unerwünschten Bereichen (z.B. $0{,}60 < p < 0{,}80$)?
- Sind schwierige und leichte Antworten in etwa gleichem Umfang vertreten (bei Power-Tests)?
- Sind die Übergänge in der Schwierigkeit zwischen den Items nicht zu groß?

Das Kriterium *Streuung*

Streuung wird auch als Differenzierungsfähigkeit bezeichnet

Die Streuung eines Tests sollte im Vergleich zu seinem Standardmessfehler groß sein, denn so lassen sich die Probanden besser in mehrere voneinander unterschiedene Gruppen unterteilen.

Setzt man eine Normalverteilung voraus, so sollten im Bereich eine Standardabweichung vom Mittelwert nach oben und eine nach unten rund 68% der Fälle liegen.

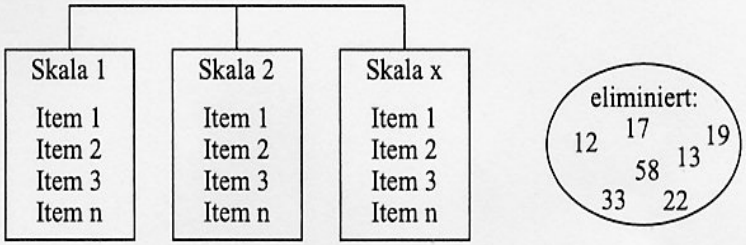
Phasen der Konstruktion von Tests

Phase 1→ Items konstruieren & vorläufiges Instrument zusammenstellen

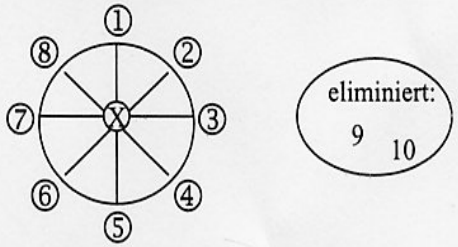
1. Bla, bla, bla... ① ② ③ ④
2. Bla, bla, bla... ① ② ③ ④
3. Bla, bla, bla... ① ② ③ ④
4. Bla, bla, bla... ① ② ③ ④

n. Bla, bla, bla... ① ② ③ ④

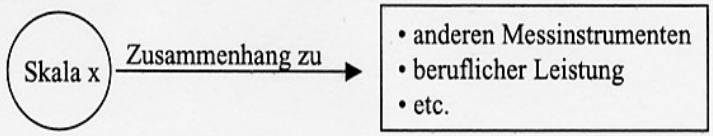
Phase 2→ Skalen bilden & unpassende Items eliminieren



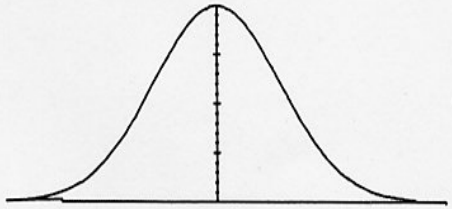
Phase 3→ Reliabilität jeder Skala berechnen & unpassende Items eliminieren



Phase 4→ Validität jeder Skala berechnen & ggf. Skalen eliminieren



Phase 5→ ggf. Normen erstellen



Itemselektion – Kurzzusammenfassung

Kurzübersicht über die (statistischen) Selektionskriterien.

Ein Item ist in der Regel (jedoch nicht ohne reifliche inhaltliche Überlegungen) zu eliminieren bei:

- einer Trennschärfe unter 0.32 (wichtigstes Kriterium)
- einer zu hohen oder zu niedrigen Schwierigkeit
- einer deutlich niedrigeren Streuung als derjenigen der anderen Items

Eine generelle Beurteilung von Schwierigkeitsindizes, Trennschärfen, Reliabilitäten und Validitäten ist schwierig, denn diese hängt ab

- vom Kontext, wie zum Beispiel der Art des verwendeten Tests (objektiver Test, Persönlichkeitstest, projektiver Test),
- der untersuchten Stichprobe (homogen/heterogen),
- der Art und
- der Breite des gemessenen Merkmals (breiter oder enger Merkmalsausschnitt)

Beurteilung der Höhe von Testkennwerten

Kennwert	Kürzel	Niedrig	Mittel	Hoch
Schwierigkeit	p	>.80	.80-.20	<.20
Trennschärfe (korrigiert)	r_{itc}	<.30	.30-.50	>.50
Objektivität (Auswerter)	r_{k}	<.60	.60-.90	>.90
Reliabilität	r_{tt}	<.80	.80-.90	>.90
Validität (unkorrigiert)	r_{tc}	<.40	.40-.60	>.60
Größe der Eichstichprobe	N	<150	150-300	>300

Eine ungefähre Richtlinie bilden dennoch die Angaben von Fisseni (1997)

Klausurteil II: Testtheorien (hier: Probeklausur zum üben)
Dr. Tobias Constantin Haupt

Aufgabe 1 (max. 2 Punkte)

Verkehrspsychologen konstruierten einen Fragebogen zur ***Attraktivität schnellen Fahrens mit dem PKW***. Der Fragebogen umfasst die im folgenden aufgeführten Items mit dichotomem Antwortmodus („ja“ = 1 und „nein“ = 0). In der folgenden Tabelle sind die Kennwerte aufgeführt, die man aus einem Pretest mit 100 Probanden erhielt.

Nr.	Item	M	r_{IT}	Z	E
1	Mich reizen nur PS-starke Autos.	0,30	0,42		
2	Wenn ich mit dem Auto fahre, denke ich an nichts anderes mehr.	0,66	0,25		
3	Am Autofahren reizt mich hauptsächlich die Geschwindigkeit.	0,55	0,47		
4	Wenn ich mit dem Auto zum Spaß rumfahre, vergeht die Zeit wie im Fluge.	0,76	0,10		
5	Wenn die Strecke frei ist, drücke ich auf's Gas.	0,58	0,50		
6	„Lahme Enten" haben auf der linken Spur nichts verloren, oder finden Sie doch?	0,20	0,15		
7	Ich habe mehr als 4 Punkte in Flensburg, finde das aber nicht so schlimm.	0,69	0,35		
8	Manchmal macht es mir Spaß, den Motor richtig hochzujagen.	0,68	0,49		
9	Schnell zu fahren macht mir Spaß.	0,45	0,57		
10	Erst wenn ich richtig Gas geben kann, fühle ich mich wohl.	0,09	0,10		

Probeklausur

Nr.	Item	M	r_{IT}	Z	E
1	Mich reizen nur PS-starke Autos.	0,30	0,42		
2	Wenn ich mit dem Auto fahre, denke ich an nichts anderes mehr.	0,66	0,25		
3	Am Autofahren reizt mich hauptsächlich die Geschwindigkeit.	0,55	0,47		
4	Wenn ich mit dem Auto zum Spaß rumfahre, vergeht die Zeit wie im Fluge.	0,76	**0,10**		X
5	Wenn die Strecke frei ist, drücke ich auf's Gas.	0,58	0,50		
6	„Lahme Enten" haben auf der linken Spur nichts verloren, **oder finden Sie doch**?	0,20	**0,15**	**unklarer Antwortbezug**	X
7	Ich habe mehr als 4 Punkte in Flensburg, finde das aber nicht so schlimm.	0,69	0,35	**2 Aussagen unklarer Antwortbezug**	X
8	Manchmal macht es mir Spaß, den Motor richtig hochzujagen.	0,68	0,49		
9	Schnell zu fahren macht mir Spaß.	0,45	0,57		
10	Erst wenn ich richtig Gas geben kann, fühle ich mich wohl.	**0,09**	**0,10**		X

(?)

Klausurteil II: Testtheorien (Oktober 2003)
Dr. Tobias Constantin Haupt

Aufgabe 1 (max. 1,5 Punkte)

Die folgenden Items sind einer Untersuchung von Studierenden der Sozialpädagogik zu ***rechtsextremen Einstellungen*** entnommen. Ziel war es, Skalen zu konstruieren, mit denen verschiedene Aspekte von Einstellungen gegenüber ***straffälligen Asylbewerbern*** erfasst werden sollten. Im Folgenden sind einige der 50 Items mit ihren Kennwerten aufgeführt, die aus einer Voruntersuchung mit N = 150 Probanden stammen. Die Personen hatten sechs abgestufte Möglichkeiten zu jedem Item Stellung zu nehmen: von „stimme sehr zu“ bis „lehne sehr ab“. Für die statistische Auswertung wurden diesen Antwortabstufungen die Zahlen ***1 bis 6*** so zugeordnet, daß ein hoher Wert für eine ***reservierte, ablehnende Haltung*** steht.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Probeklausur

Nr.	Item	M	s	r_{IT}	Anmerkungen?	E
1	Der Staat sollte mehr Geld für den Schutz seiner Bürger vor kriminellen Asylbewerbern ausgeben.	3,63	1,51	.22		
2	Weil viele Leute straffälligen Asylbewerbern mit Vorurteilen begegnen, treiben sie sie noch mehr in die Straffälligkeit.	2,18	1,34	.27		
3	Es lohnt sich, straffällige Asylbewerber zu „erziehen“, denn wer einmal straffällig wird, der wird es immer wieder.	1,91	1,13	.52	**sprachlich unlogisch → Logik**	**X**
4	Man sollte straffälligen Asylbewerbern mit Mißtrauen begegnen.	2,47	1,30	.25		
5	Es gibt geborene Verbrecher, deshalb gehören asylsuchende Straftäter am besten gleich weggesperrt und nach Hause geschickt.	3,04	1,81	.34		
6	Eine kräftige Tracht Prügel wäre oft ganz gut für asylsuchende Straftäter.	3,08	1,80	.31		
7	Kriminelle Asylbewerber sind häufig geisteskrank.	2,39	1,27	.02		
8	Würden Sie Bemühungen unterstützen, spezielle Gefängnisse für straffällige Asylbewerber abzuschaffen oder nicht?	2,03	1,70	.20		
9	Auch wenn ein Asylbewerber schon einmal verurteilt worden ist, so muß er im Alltag trotzdem genauso behandelt werden wie ein nicht straffälliger Asylbewerber.	1,71	1,10	.35		
10	Straffällige Asylbewerber nehmen oft Drogen und sind Tagediebe.	1,80	1,21	.30		
11	Gefängnisse sozialisieren die asylsuchenden Straftäter noch zusätzlich für eine „Kriminellenkarriere“.	3,80	1,59	.33		
12	Körperliche Züchtigung tut Not bei asylsuchenden Straftätern, damit sie mal kapieren, wo es langgeht in diesem Land.	2,84	1,37	.31		
13	Asylsuchende Straftäter kommen oft aus zerrütteten Familien.	1,95	0,56	.17		

(?)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Probeklausur

Nr.	Item	M	s	r_{IT}	Anmerkungen?	E
1	Der Staat sollte mehr Geld für den Schutz seiner Bürger vor kriminellen Asylbewerbern ausgeben.	3,63	1,51	.22		
2	Weil viele Leute straffälligen Asylbewerbern mit Vorurteilen begegnen, treiben sie sie noch mehr in die Straffälligkeit.	2,18	1,34	.27		
3	Es lohnt sich, straffällige Asylbewerber zu „erziehen“, denn wer einmal straffällig wird, der wird es immer wieder.	1,91	1,13	.52	sprachlich unlogisch → Logik	X
4	Man sollte straffälligen Asylbewerbern mit Mißtrauen begegnen.	2,47	1,30	.25	zur Probe drinlassen	
5	Es gibt geborene Verbrecher, deshalb gehören asylsuchende Straftäter am besten gleich weggesperrt und nach Hause geschickt.	3,04	1,81	.34	mind. 2 sprachliche Inhalte	X
6	Eine kräftige Tracht Prügel wäre oft ganz gut für asylsuchende Straftäter.	3,08	1,80	.31		
7	Kriminelle Asylbewerber sind häufig geisteskrank.	2,39	1,27	.02		X
8	Würden Sie Bemühungen unterstützen, spezielle Gefängnisse für straffällige Asylbewerber abzuschaffen oder nicht?	2,03	1,70	.20	unklarer Antwortbezug	X
9	Auch wenn ein Asylbewerber schon einmal verurteilt worden ist, so muß er im Alltag trotzdem genauso behandelt werden wie ein nicht straffälliger Asylbewerber.	1,71	1,10	.35		
10	Straffällige Asylbewerber nehmen oft Drogen und sind Tagediebe.	1,80	1,21	.30	- 2 Inhalte - oft (?)	X
11	Gefängnisse sozialisieren die asylsuchenden Straftäter noch zusätzlich für eine „Kriminellenkarriere“.	3,80	1,59	.33		
12	Körperliche Züchtigung tut Not bei asylsuchenden Straftätern, damit sie mal kapieren, wo es langgeht in diesem Land.	2,84	1,37	.31	- 2 Inhalte - ethisch?	X
13	Asylsuchende Straftäter kommen oft aus zerrütteten Familien.	1,95	0,56	.17		X

(?)

Aufgabe 4 (max. 2,5 Punkte)

Ein Verkehrspsychologe verwendet in einer Untersuchung eine kurze Skala zur ***Attraktivität riskanten Fahrens mit dem Motorrad***. Diese Skala umfaßt 4 Items. Das Antwortformat ist eine fünfstufige Ratingskala. Hohe Zahlen bedeuten hohe Merkmalsausprägungen. Zusätzlich wurde erfaßt, ob und ggf. wie viele Unfälle die Versuchsperson im letzten Kalenderjahr mit dem Motorrad hatte (Spalte „Unfälle"). Es ergaben sich die folgenden Daten:

Person	Item 1	Item 2	Item 3	Item 4	Unfälle				
1	3	1	1	2	0				
2	3	2	2	2	0				
3	3	2	3	2	0				
4	5	5	5	5	2				
5	3	1	2	2	0				
6	4	5	5	5	1				
7	1	3	1	4	0				
8	5	3	3	4	1				

Aufgabe 4 (max. 2,5 Punkte)

Ein Verkehrspsychologe verwendet in einer Untersuchung eine kurze Skala zur ***Attraktivität riskanten Fahrens mit dem Motorrad***. Diese Skala umfaßt 4 Items. Das Antwortformat ist eine fünfstufige Ratingskala. Hohe Zahlen bedeuten hohe Merkmalsausprägungen. Zusätzlich wurde erfaßt, ob und ggf. wie viele Unfälle die Versuchsperson im letzten Kalenderjahr mit dem Motorrad hatte (Spalte „Unfälle"). Es ergaben sich die folgenden Daten:

Person	Item 1	Item 2	Item 3	Item 4	Unfälle	∑ 1-4	∑ ¬ 2		
1	3	1	1	2	0	**7**	**6**		
2	3	2	2	2	0	**9**	**7**		
3	3	2	3	2	0	**10**	**8**		
4	5	5	5	5	2	**20**	**15**		
5	3	1	2	2	0	**8**	**7**		
6	4	5	5	5	1	**19**	**14**		
7	1	3	1	4	0	**9**	**6**		
8	5	3	3	4	1	**15**	**12**		

Berechnen Sie die nachfolgenden Größen und tragen Sie sie in die Tabelle ein:

a) Korr. Trennschärfe für Item 2	$r_{IT} = 0{,}87$
b) Testhalbierungsreliabilität (odd-even)	$r_{odd\text{-}even}$: .67 $R = \frac{2 \cdot 0{,}67}{1{,}67} = \underline{\underline{0{,}8}}$
c) Interne Konsistenz	$\alpha = \frac{4}{3} \cdot \left(1 - \frac{1{,}696 + 2{,}5 + 2{,}5 + 1{,}929}{26{,}41} \right) =$ $= \frac{4}{3} \cdot \left(\frac{1 - 8{,}625}{26{,}41} \right) = \underline{\underline{0{,}90}}$
d) Validität der Skala hinsichtlich des Kriteriums Unfallhäufigkeit	$r_{tc} = 0{,}94$
e) Sie möchten für den hypothetischen Fall einer absoluten Reliabilität der Skala wissen, wie hoch die maximale Kriteriumsvalidität dann wäre. Was müssen sie für die Beantwortung dieser Frage berechnen? Begründen Sie ihre Antwort stichwortartig. Die Berechung selbst brauchen sie NICHT durchzuführen.	

Zur **Herleitung des klassischen testtheoretischen Modells** werden die folgenden fünf (man kann sie auch zu drei oder vier zusammenfassen) Axiome benötigt. Dabei handelt es sich um Festsetzungen, bzw. Definitionen, deren empirische Adäquatheit zunächst unbewiesen bleibt.

Axiome sind also Grundannahmen.

Ein fixer Gesamtüberblick über die fünf Axiome (kurz & schmerzlos):

- $X = T + e$
- $\mu(e) = 0$
- $\rho_{(T,e)} = 0$
- $\rho_{(ex,Ty)} = 0$
- $\rho_{(ex,\, ey)} = 0$

1. Axiom: $x_j = w_j + e_j$

- **j = Index einer Person j**
- **x_j = gemessener Wert einer Person j**
- **w_j = der wahre Wert einer Person j**
- **e_j = Fehlerwert einer Person j**

Oder anders ausgedrückt: „Wer misst, misst Mist.“ – zumindest teilweise!

Beispiel: Das beobachtete Intelligenztestergebnis einer Person setzt sich zusammen aus ihrer „wahren“ Intelligenz und (Mess-)Fehlereffekten (z.B. wegen Müdigkeit, Unkonzentriertheit).

Das Konzept des Messfehlers: Messfehler umfassen die Gesamtheit aller unsystematischen (!) und nicht kontrollierbaren oder vorhersagbaren potentiellen Einflußgrößen auf das Messergebnis.

2. Axiom

Der (bei häufiger Messwiederholung) erwartete Mittelwert (μ) der Messfehler ist Null: **μ(E) = 0**

D.h., daß es bei wiederholten Testanwendungen unter identischen Bedingungen zu einem Fehlerausgleich (Ausmittelung von Fehlerschwankungen) kommt und der gemittelte Testwert bei einer Person über alle Messungen dem wahren Wert nahezu entspricht.

3. Axiom

Die Höhe des Messfehlers E ist unabhängig vom (wahren) Ausprägungsgrad T des getesteten Merkmals, d.h., wahrer Wert und Fehlerwert sind unkorreliert: $\mathbf{r_{TE} = 0}$.

Beispiel: Fehlereinflüsse durch die Tagesform sind bei Personen mit hoher und niedriger Intelligenz in gleicher Weise wirksam.

Beispiel

aus Rost (1996), S. 41, Aufg. 1

Sie haben:
- die Meßwerte einer Variablen x
- die wahren Werte derselben Variable T_X
- von 5 Personen

Tabellarisch sieht das z.B. so aus:

Pbn	Meßwert x	wahrer Wert T_X
1	2	1
2	0	1
3	5	5
4	10	9
5	8	9

- Prüfen sie, ob für den Meßfehler der Meßwerte x die beiden Axiome 2 und 3 der KTT (= Meßfehlertheorie) gelten.
- Berechnen Sie die Reliabilität!

Beispiel

aus Rost (1996), S. 41, Aufg. 1

Sie haben:
- die Meßwerte einer Variablen x
- die wahren Werte derselben Variable T_X
- von 5 Personen

Tabellarisch sieht das z.B. so aus:

Pbn	Meßwert x	wahrer Wert T_X	Meßfehler
1	2	1	1
2	0	1	-1
3	5	5	0
4	10	9	1
5	8	9	-1

- Prüfen sie, ob für den Meßfehler der Meßwerte x die beiden Axiome 2 und 3 der KTT (= Meßfehlertheorie) gelten.
- Berechnen Sie die Reliabilität!

2. Axiom ✓
$\mu(E) = 0$

3. Axiom
$r_{TE} = 0$