Grundlagen der Testtheorie WS 2020/21

10. Reliabilität

25.01.2021

Prof. Dr. Eunike Wetzel

Semesterplan

Sitzung	Termin	Thema
1	02.11.	Grundlagen & Gütekriterien
2	09.11.	Schritte der Testkonstruktion: Übersicht Konstruktdefinition & Itemgenerierung
3	16.11.	Erstellung eines Testentwurfs
4	23.11.	Klassische Testtheorie
5	07.12.	Item Response Theorie
6	14.12.	Exploratorische Faktorenanalyse
7	04.01.	Itemanalyse 1
8	11.01.	Itemanalyse 2, Itemselektion & Testrevision
9	18.01.	Objektivität
10	**25.01.**	**Reliabilität**
11	01.02.	Validität
12	08.02.	Normierung, Standards für psychologisches Testen

Reliabilität

1. Definition Reliabilität
2. Methoden zur Reliabilitätsschätzung
3. Einflussfaktoren auf die Höhe der Reliabilität
4. Anwendung: Konfidenzintervalle in der Individualdiagnostik

1. Definition Reliabilität

- Die Reliabilität gibt den Grad der Messgenauigkeit eines Testwerts an
- Objektivität ist Voraussetzung für eine hohe Reliabilität
- Die Reliabilität liegt zwischen 0 und 1

$$\mathrm{Rel}(\mathrm{X}_i) = \frac{Var(\tau_i)}{Var(\mathrm{X}_i)} = \frac{Var(\tau_i)}{Var(\tau_i) + Var(\varepsilon_i)}$$

- In der Praxis kann die Reliabilität nicht exakt berechnet werden, daher wird sie mit verschiedenen Methoden geschätzt

Reliabilität

1. Definition Reliabilität
2. **Methoden zur Reliabilitätsschätzung**
    1. Retest-Reliabilität
    2. Paralleltest-Reliabilität
    3. Testhalbierungsreliabilität
    4. Interne Konsistenz
3. Einflussfaktoren auf die Höhe der Reliabilität
4. Anwendung: Konfidenzintervalle in der Individualdiagnostik

2. Methoden zur Reliabilitätsschätzung

- Die meisten Methoden zur Reliabilitätsschätzung basieren statistisch gesehen auf Korrelationen zwischen…
  - Testwerten aus einem Test zu zwei Messzeitpunkten
  - Testwerten aus zwei parallelen Tests, die direkt hintereinander erhoben werden
  - Testwerten aus einem Test, der aus mindestens zwei Testteilen besteht

2.1 Retest-Reliabilität

- Test wird an der gleichen Stichprobe zu zwei verschiedenen Zeitpunkten durchgeführt und die Korrelation der Testwerte $r_{x1,x2}$ berechnet
- Annahmen:
  - Konstante wahre Werte
  - Konstante Messfehlereinflüsse
- Unter diesen Annahmen entspricht die Korrelation der Testwerte dem Anteil der wahren Varianz an der Varianz der Testwerte

2.1 Retest-Reliabilität

$$
\begin{aligned}
\mathrm{Corr}(x_1, x_2) &= \frac{\mathrm{Cov}(x_1, x_2)}{\mathrm{SD}(x_1) \cdot \mathrm{SD}(x_2)} \\
&= \frac{\mathrm{Cov}(\tau_1 + \varepsilon_1, \tau_2 + \varepsilon_2)}{\mathrm{SD}(x_1) \cdot \mathrm{SD}(x_2)} \\
&= \frac{\mathrm{Cov}(\tau_1, \tau_2)}{\mathrm{SD}(x_1) \cdot \mathrm{SD}(x_2)} \\
&= \frac{\mathrm{Var}(\tau)}{\mathrm{Var}(x)} \\
&= \mathrm{Rel}(x)
\end{aligned}
$$

Moosbrugger & Kelava (2012), S. 122

2.1 Retest-Reliabilität

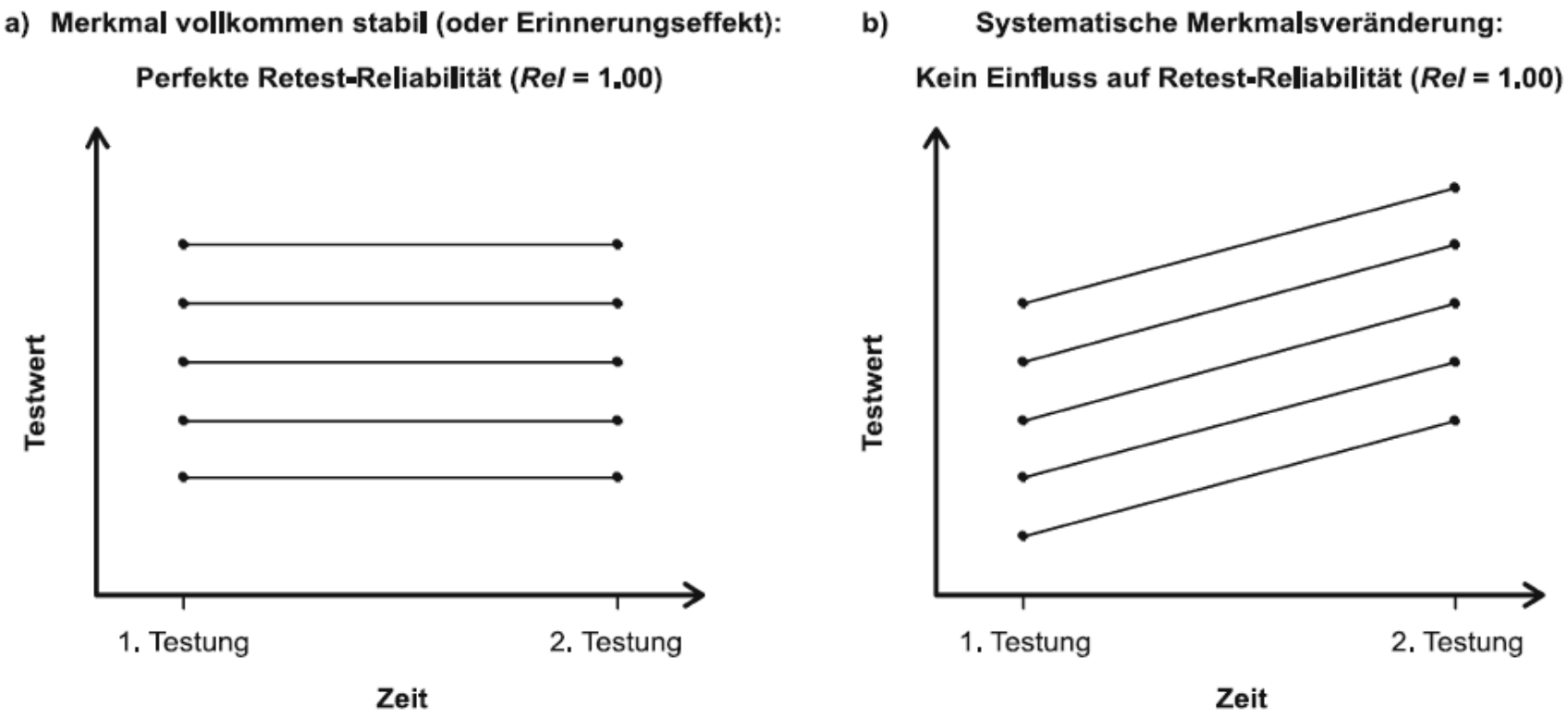


Abb. 6.1 Moosbrugger & Kelava (2012)

2.1 Retest-Reliabilität

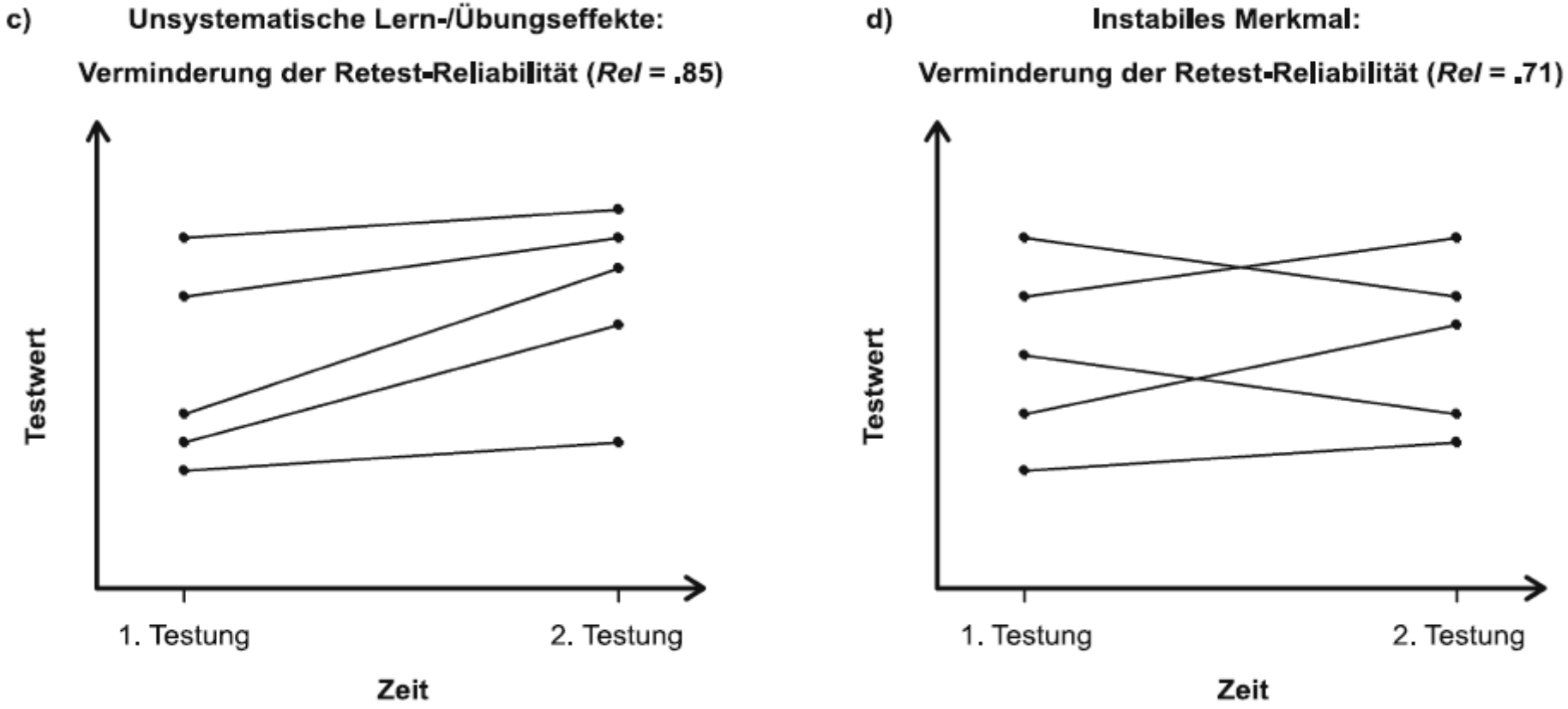


Abb. 6.1 Moosbrugger & Kelava (2012)

2.1 Retest-Reliabilität

Tabelle 25: Retest-Reliabilitäten der NEO-PI-R-Form S Skalen (1 Monat–2 Jahre)

NEO-PI-R-Skalen, Form S	Test-Retest-Intervall in Monaten				
	1 (70)	2 (119)	6 (28)	12 (11)	24 (10)
Hauptskalen					
Neurotizismus	.91	.90	.91	.93	.97
Extraversion	.91	.88	.87	.82	.95
Offenheit für Erfahrungen	.89	.82	.84	.94	.90
Verträglichkeit	.88	.88	.84	.80	.85
Gewissenhaftigkeit	.91	.90	.94	.87	.90
Facetten					
Neurotizismus					
Ängstlichkeit	.86	.83	.92	.72	.94
Reizbarkeit	.73	.85	.79	.95	.57
Depression	.86	.80	.86	.81	.91
Soziale Befangenheit	.75	.86	.83	.87	.92
Impulsivität	.79	.74	.84	.81	.69
Verletzlichkeit	.88	.86	.82	.87	.85

2.2 Paralleltest-Reliabilität

- Parallele Testformen werden derselben Stichprobe vorgegeben und die resultierenden Testwerte korreliert

$$\mathrm{Rel}(x) = \mathrm{Corr}(x_A, x_B)$$

- Wird häufig bei Leistungstests eingesetzt
- Beispiel Zahlenreihe:
  2 5 8 11 ?
  4 8 12 16 ?
- Bei parallelen Testformen können auch Übungs- bzw. Transfereffekte auftreten
- Die Parallelität von Testformen kann mithilfe der konfirmatorischen Faktorenanalyse überprüft werden

2.3 Testhalbierungsreliabilität

- Auch Split-half Reliabilität genannt
- Berechnung der Korrelation von Testwerten aus zwei Testhälften
- Vorgehen:
  1. Items des Tests werden in zwei möglichst parallele Testhälften aufgeteilt
  2. Korrelation der Testwerte aus beiden Testhälften (Halbtest-Reliabilität)
  3. Aufwertung der Halbtest-Reliabilität mit der Spearman-Brown-Formel

2.3 Testhalbierungsreliabilität

Spearman-Brown-Formel

$$\mathrm{Rel}\left(x_{\text{vollständig}}\right) = \frac{2\mathrm{Corr}\left(x_p, x_q\right)}{1 + \mathrm{Corr}\left(x_p, x_q\right)} = \frac{2\mathrm{Rel}\left(x_{\text{halb}}\right)}{1 + \mathrm{Rel}\left(x_{\text{halb}}\right)}$$

- Mit der Spearman-Brown-Formel lässt sich auch berechnen, um wie viele parallele Items ein bestehender Test verlängert werden muss, um eine bestimmte Reliabilität zu erreichen

2.3 Testhalbierungsreliabilität

Methoden zur Aufteilung der Items auf 2 Testteile:

1. **Odd-Even**: Ungerade Items in eine Hälfte, gerade Items in die andere Hälfte
   sinnvoll bei Leistungstests mit Items, die in ihrer Schwierigkeit ansteigen
2. **Zeitpartitionierungsmethode**: Items werden nach der Bearbeitungszeit aufgeteilt
   sinnvoll bei Tests mit vielen gleichartigen Items
3. **Itemzwillinge**: Bildung von Itempaaren anhand von Schwierigkeit und Trennschärfe, zufällige Zuweisung zu einer Testhälfte
   sinnvoll bei heterogenen Items
4. **Zufällige Aufteilung**

2.3 Testhalbierungsreliabilität

- Vorteil: Übungs- und Ermüdungseffekte verteilen sich gleichmäßig auf die beiden Testhälfte (außer bei Zeitpartitionierung)
- Nachteile:
  - Keine Garantie, dass die gebildeten Testhälften tatsächlich parallel sind
  - Bei nicht perfekt parallelen Testhälften kann es zu einer Unterschätzung der Reliabilität des Gesamttests kommen
  - Aufteilungsmethoden können zu unterschiedlichen Reliabilitätsschätzungen führen

Bsp. NARQ

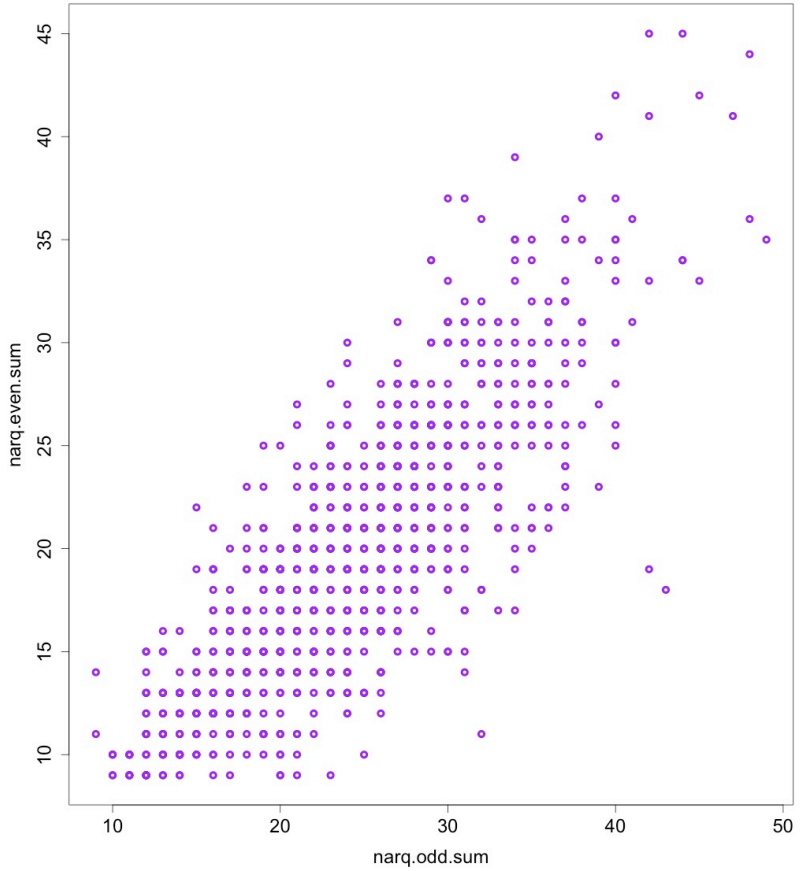
Odd-Even Aufteilung

```
> narq.odd <- subset(narq, select=c("narq1","narq3","narq5","narq7","narq9","narq11","narq13",
+ "narq15","narq17"))
> narq.even <- subset(narq, select=c("narq2","narq4","narq6","narq8","narq10","narq12","narq14",
+ "narq16","narq18"))
```

$$r_{odd,even} = 0.81$$

$$\text{Rel}_{vollständig} = \frac{2 \cdot r_{odd,even}}{1 + r_{odd,even}} = \frac{2 \cdot 0.81}{1 + 0.81}$$

$$= 0.90$$



Bsp. NARQ

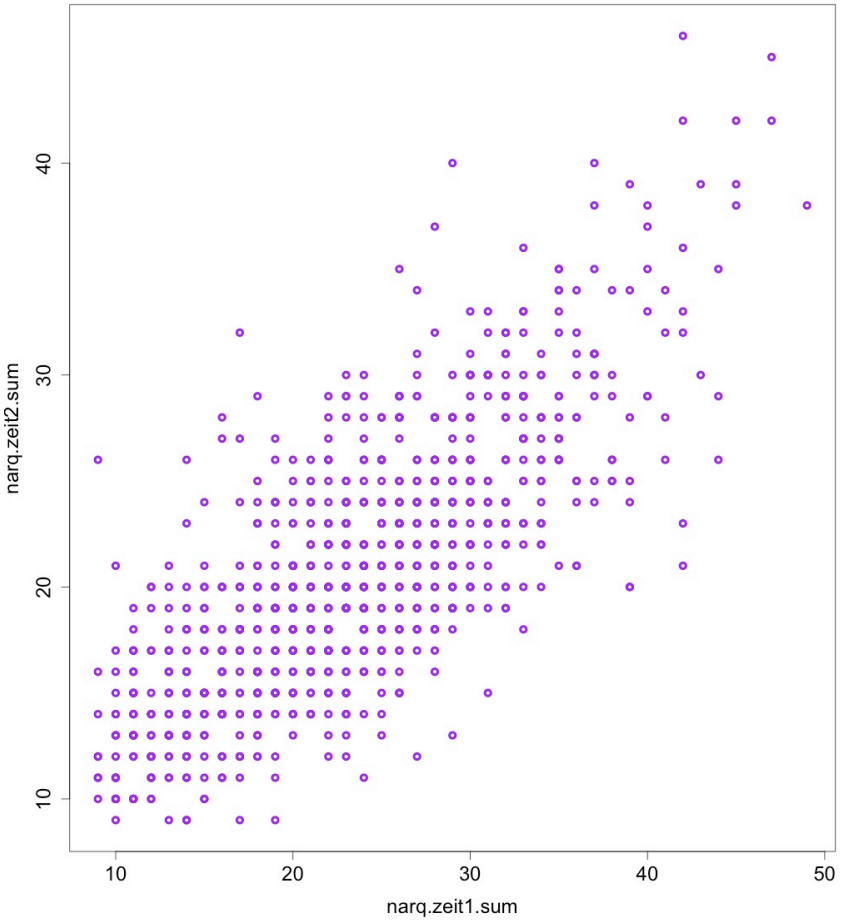
Zeitpartitionierung

```
> narq.zeit1 <- subset(narq, select=narq1:narq9)
> narq.zeit2 <- subset(narq, select=narq10:narq18)
```

$$r_{zeit1,zeit2} = 0.77$$

$$\text{Rel}_{vollständig} = \frac{2 \cdot r_{zeit1,zeit2}}{1 + r_{zeit1,zeit2}} = \frac{2 \cdot 0.77}{1 + 0.77}$$

$$= 0.87$$



Bsp. NARQ

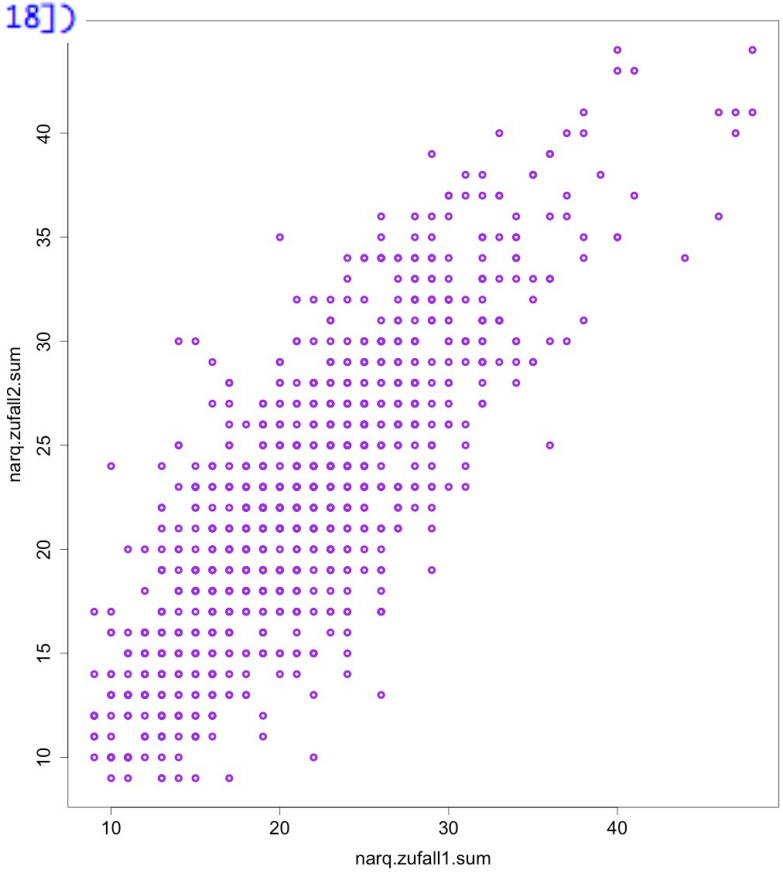
Zufällige Aufteilung

```
> zufallsfolge <- sample(seq(1:18))
> zufallsfolge
 [1] 13  2  9 10  1  4  5 15 14  6  8 12  3 18 16  7 17 11
> narq.zufall1 <- subset(narq, select=zufallsfolge[1:9])
> narq.zufall2 <- subset(narq, select=zufallsfolge[10:18])
```

$$r_{zufall1,zufall2} = 0.82$$

$$\text{Rel}_{vollständig} = \frac{2 \cdot r_{zufall1,zufall2}}{1 + r_{zufall1,zufall2}} = \frac{2 \cdot 0.82}{1 + 0.82}$$

$$= 0.90$$



2.4 Interne Konsistenz

- Auch innere Konsistenz genannt
- Verallgemeinerung der Testhalbierungsreliabilität auf beliebig viele Testteile
- In der Regel werden *m* Testteile verwendet, wobei *m* die Anzahl der Items darstellt
- Der am häufigsten verwendete Koeffizient zur Bestimmung der internen Konsistenz ist Cronbachs alpha (Cronbach, 1951)

2.4 Interne Konsistenz

Cronbachs alpha

$$\mathrm{Rel}(x) = \alpha = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{i=1}^{m} Var(x_i)}{Var(x)} \right)$$

m = Anzahl der Items

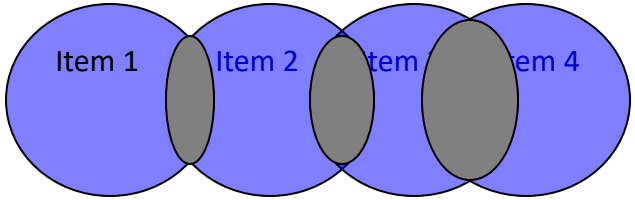
$Var(x_i)$ = Varianz des *i*-ten Items

Var(x) = Varianz des Gesamttests x

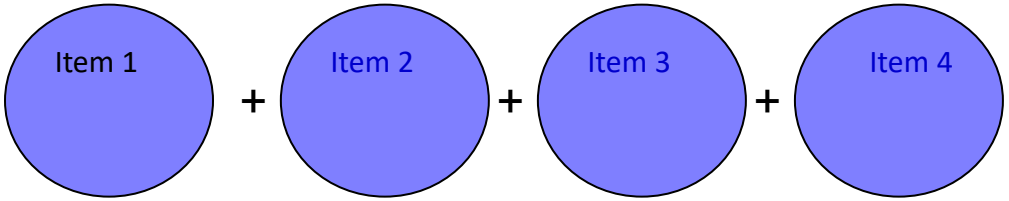
2.4 Interne Konsistenz

Grafische Veranschaulichung

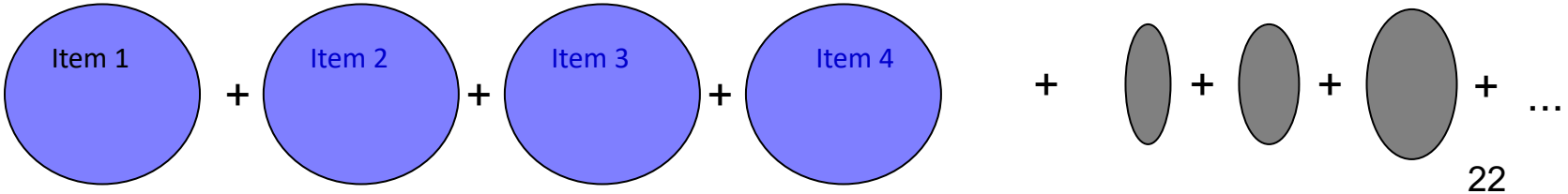
4 Items mit Varianzen (blau) und Kovarianzen (grau)



Im Zähler steht die Summe der Itemvarianzen



Im Nenner steht die Summe der Itemvarianzen und der Kovarianzen



2.4 Interne Konsistenz

Berechnung der Varianz eines Summenscores X

X = X1 + X2

$$Var_X = \begin{pmatrix} Var_{X1} & Cov_{X1X2} \\ Cov_{X1X2} & Var_{X1} \end{pmatrix}$$

X = X1 + X2 + X3

$$Var_X = \begin{pmatrix} Var_{X1} & Cov_{X1X2} & Cov_{X1X3} \\ Cov_{X1X2} & Var_{X2} & Cov_{X2X3} \\ Cov_{X1X3} & Cov_{X2X3} & Var_{X3} \end{pmatrix}$$

2.4 Interne Konsistenz

Cronbachs alpha

$$\mathrm{Rel}(x) = \alpha = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{i=1}^{m} Var(x_i)}{Var(x)} \right)$$

Var(x) für 2 Items:

$$Var(x) = Var(x_1 + x_2) = Var(x_1) + Var(x_2) + 2 \cdot Cov(x_1, x_2)$$

→ Je stärker die Items positiv korreliert sind, desto größer wird Cronbachs alpha

Bsp. NARQ

Item	Varianz
narq1	1.95
narq2	1.53
...	...
narq18	1.70
Σ	29.89

Summenscore $narq_{sum}$
Varianz = 172.39

$$\text{Rel}(x) = \alpha = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{i=1}^{m} Var(x_i)}{Var(x)} \right) = \frac{18}{17} \cdot \left(1 - \frac{29.89}{172.39} \right) = 0.88$$

2.4 Interne Konsistenz

Cronbachs alpha & Anzahl der Items

- Cronbachs alpha steigt mit der Itemzahl, allerdings nur bei positiven Korrelationen zwischen den Items
- Negative Korrelationen zwischen Items können Cronbachs alpha reduzieren

$$Var(x) = Var(x_1 + x_2) = Var(x_1) + Var(x_2) + 2 \cdot Cov(x_1, x_2)$$

2.4 Interne Konsistenz

Cronbachs alpha

- Die korrekte Schätzung der Reliabilität durch Cronbachs alpha setzt essenziell tau-äquivalente Items voraus
- Sind die Items nicht essenziell tau-äquivalent, ist Cronbachs alpha eine untere Schranke für die Reliabilität (sofern die Annahme unkorrelierter Fehler erfüllt ist)
- Für tau-kongenerische Items gibt es andere Koeffizienten zur Schätzung der Reliabilität, z.B. McDonalds Omega
- Bei heterogenen (gering korrelierten) Items kann die Reliabilität durch interne Konsistenzmaße deutlich unterschätzt werden

2.4 Interne Konsistenz

Aspekte der Interpretation von Cronbachs alpha

- Eindimensionalität
- Negativ gepolte Items
- Negatives Cronbachs alpha

2.4 Interne Konsistenz

Cronbachs alpha ist kein Maß für Eindimensionalität

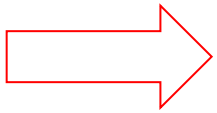
$$R = \begin{pmatrix} 1 & 0{,}8 & 0{,}8 & 0{,}8 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}8 & 1 & 0{,}8 & 0{,}8 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}8 & 0{,}8 & 1 & 0{,}8 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}8 & 0{,}8 & 0{,}8 & 1 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 1 & 0{,}8 & 0{,}8 & 0{,}8 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}8 & 1 & 0{,}8 & 0{,}8 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}8 & 0{,}8 & 1 & 0{,}8 \\ 0{,}4 & 0{,}4 & 0{,}4 & 0{,}4 & 0{,}8 & 0{,}8 & 0{,}8 & 1 \end{pmatrix}$$

- $\alpha = .91$

2.4 Interne Konsistenz

- Ergebnis der Faktorenanalyse

$$R = \begin{pmatrix} 1 & 0,8 & 0,8 & 0,8 & 0,4 & 0,4 & 0,4 & 0,4 \\ 0,8 & 1 & 0,8 & 0,8 & 0,4 & 0,4 & 0,4 & 0,4 \\ 0,8 & 0,8 & 1 & 0,8 & 0,4 & 0,4 & 0,4 & 0,4 \\ 0,8 & 0,8 & 0,8 & 1 & 0,4 & 0,4 & 0,4 & 0,4 \\ 0,4 & 0,4 & 0,4 & 0,4 & 1 & 0,8 & 0,8 & 0,8 \\ 0,4 & 0,4 & 0,4 & 0,4 & 0,8 & 1 & 0,8 & 0,8 \\ 0,4 & 0,4 & 0,4 & 0,4 & 0,8 & 0,8 & 1 & 0,8 \\ 0,4 & 0,4 & 0,4 & 0,4 & 0,8 & 0,8 & 0,8 & 1 \end{pmatrix}$$



Rotierte Faktorenmatrix(a)

	Faktor	
	1	2
v1	.232	.864
v2	.232	.864
v3	.232	.864
v4	.232	.864
v5	.864	.232
v6	.864	.232
v7	.864	.232
v8	.864	.232

- α = .91
- → Die interne Konsistenz kann auch dann hoch sein, wenn die Items ein mehrdimensionales Konstrukt messen

2.4 Interne Konsistenz

Negativ gepolte Items

- Wenn negativ gepolte Items einen eigenen Faktor bilden, können Maße der internen Konsistenz die tatsächliche Reliabilität unter- oder überschätzen
- Wurden negativ gepolte Items nicht rekodiert, kann Cronbachs alpha negativ werden

Negatives Cronbachs alpha

- Bei negativen Korrelationen zwischen Items kann Cronbachs alpha einen negativen Wert annehmen

2.4 Interne Konsistenz

Tabelle 24: Interne Konsistenz der deutschen NEO-PI-R-Form S Facettenskalen

NEO-PI-R-Skalen	USA[a]	Alters- und Geschlechtsgruppen (deutsch								
		Allgemeine Bevölkerung								
		≥16			16–20		21–24		25–29	
		G	m	w	m	w	m	w	m	w
N =	1539	11724	4219	7505	480	1686	1358	1925	943	1189
Neurotizismus										
Ängstlichkeit	.78	.82	.79	.82	.74	.79	.79	.81	.80	.83
Reizbarkeit	.75	.73	.72	.73	.71	.72	.72	.74	.72	.74
Depression	.81	.85	.84	.85	.82	.84	.84	.85	.84	.85
Befangenheit	.68	.72	.70	.72	.65	.70	.71	.74	.73	.74
Impulsivität	.70	.64	.61	.64	.56	.63	.62	.62	.60	.61
Verletzlichkeit	.77	.79	.79	.78	.76	.78	.80	.78	.80	.78
Extraversion										
Herzlichkeit	.73	.71	.72	.69	.73	.71	.74	.70	.70	.70
Geselligkeit	.72	.80	.79	.78	.81	.79	.80	.79	.78	.79
Durchsetzungsfähigkeit	.77	.80	.80	.80	.79	.81	.80	.82	.80	.79
Aktivität	.63	.70	.70	.69	.65	.68	.69	.71	.71	.69
Erlebnishunger	.65	.60	.59	.60	.54	.56	.54	.53	.53	.53
Frohsinn	.73	.80	.79	.81	.77	.80	.79	.80	.80	.82

Übersicht

Vor- und Nachteile der Methoden zur Reliabilitätsschätzung

	Retest	Parallel-test	Split-half	Interne Konsistenz
Parallelform notwendig	nein	ja	nein	nein
Mehrere Items notwendig	nein	ja	ja	ja
2 Testdurchführungen notwendig	ja	ja	nein	nein
2 Messzeitpunkte notwendig	ja	nein	nein	nein
Überschätzung bei Erinnerungseffekten	ja	nein	nein	nein
Unterschätzung bei unsystematischer Merkmalsveränderung	ja	nein	nein	nein
Unterschätzung bei heterogenen Items	nein	nein	ja	ja

Reliabilität

1. Definition Reliabilität
2. Methoden zur Reliabilitätsschätzung
    1. Retest-Reliabilität
    2. Paralleltest-Reliabilität
    3. Testhalbierungsreliabilität
    4. Interne Konsistenz
3. **Einflussfaktoren auf die Höhe der Reliabilität**
4. Anwendung: Konfidenzintervalle in der Individualdiagnostik

3.1 Homogenität oder Heterogenität der Items

Fragebogen zur Erfassung des Organisationsklimas (Daumenlang, Müskens & Harder, 2004)

Skala Bewertung der Arbeit, Cronbachs alpha = .90

1. Wie ich meine Arbeit erlebe

	stimmt vollkommen	stimmt weitgehend	stimmt eher	stimmt eher nicht	stimmt kaum	stimmt gar nicht
1. Meine Arbeit ist interessant	○	○	○	○	○	○
2. Ich bin stolz auf meine Arbeit	○	○	○	○	○	○
3. Ich bin entsprechend meiner Fähigkeiten eingesetzt	○	○	○	○	○	○
4. Meine Arbeit ist sinnvoll	○	○	○	○	○	○
5. Ich habe eine Arbeit mit großer Verantwortung	○	○	○	○	○	○
6. Mir gefällt meine Arbeit	○	○	○	○	○	○

3.1 Homogenität oder Heterogenität der Items

Wiener Entwicklungstest (Kastner-Koller & Deimann, 2002)

- 13 verschiedene Subtests zur Erfassung des Entwicklungsstandes des Kindes
- Subtest *Turnen*
  Beispielitem: einbeiniges, freihändiges Stehen mit geschlossenen Augen für mindestens 3 sek.
- Subtest *Puppenspiel*: mit dem Spielmaterial sollen vorgesprochene Sätze dargestellt werden
  Beispielitem 1: „Der Vater streichelt den Hund."
  Beispielitem 2: „Der Hund beißt den Vater, der das Mädchen festhält."
- Cronbachs alpha für Subtests: .66 - .90

3.1 Homogenität oder Heterogenität der Items

- Tests mit homogenen Items haben meistens eine hohe Reliabilität, da die Items sehr ähnlich sind und daher hoch positiv miteinander korrelieren
- Bei Tests mit heterogenen Items kann die Reliabilität durch Maße der internen Konsistenz unterschätzt werden
- Items so zu selektieren, dass die Reliabilität möglichst hoch wird, kann die Konstruktvalidität beeinträchtigen

3.2 Testlänge

Die Reliabilität eines Tests lässt sich durch die Hinzunahme paralleler Testteile steigern

Spearman-Brown-Formel:

- Berechnung der neuen Reliabilität Rel* bei Verlängerung des Tests um den Faktor *k*

$$\mathrm{Rel}_k^* = \frac{k \cdot \mathrm{Rel}}{1 + (k-1) \cdot \mathrm{Rel}}$$

- Berechnung des Faktors *k*, um den man einen Test verlängern muss, um eine Reliabilität von Rel* zu erreichen

$$k = \frac{\mathrm{Rel}^* \cdot (1\text{-}\mathrm{Rel})}{\mathrm{Rel} \cdot (1 - \mathrm{Rel}^*)}$$

3.2 Testlänge

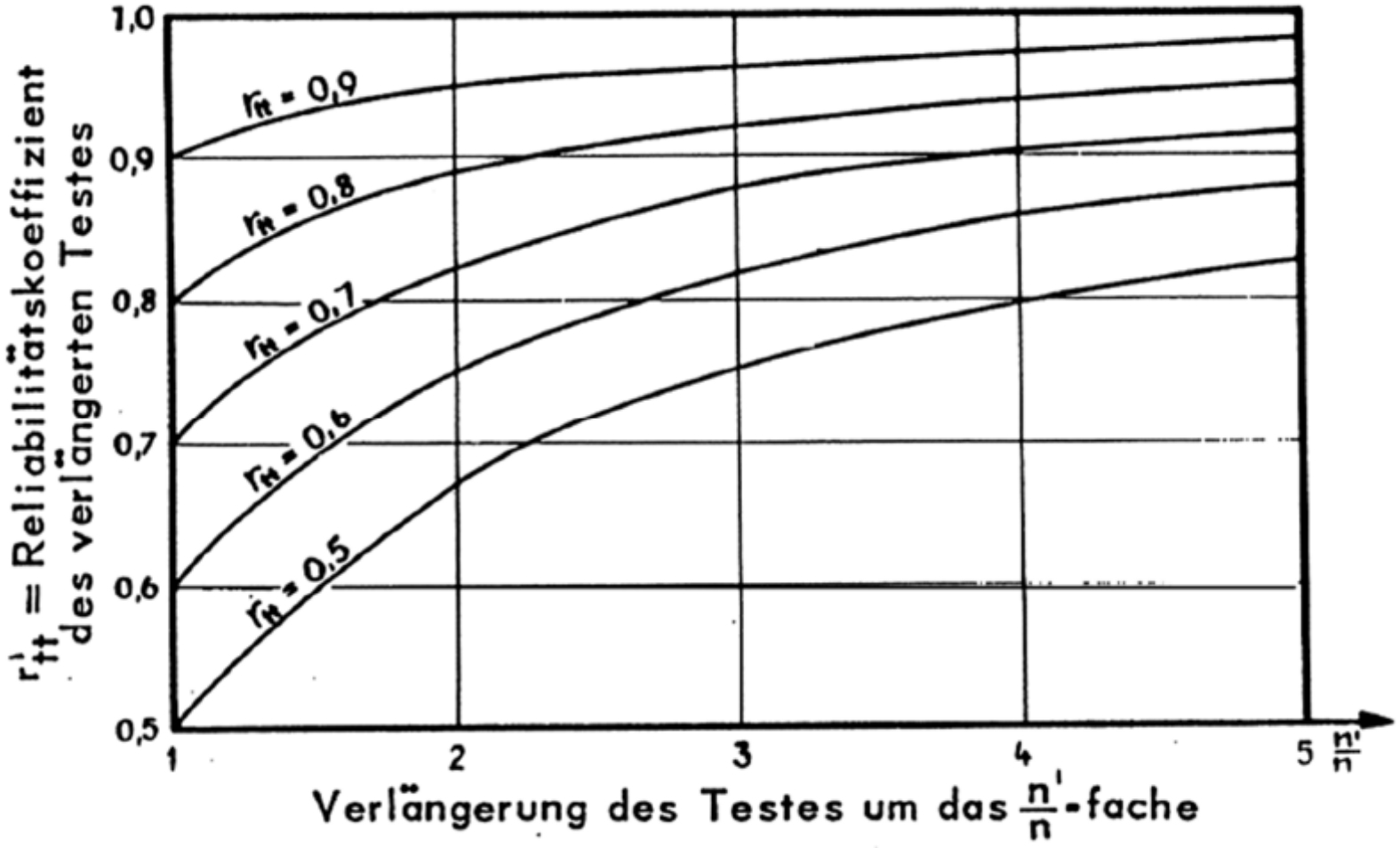


Abb. aus Lienert (1989)

Bsp. NARQ

Ziel sei es, eine Reliabilität von .95 zu erreichen

Um welchen Faktor müsste der NARQ (18 Items) verlängert werden, um dies zu erreichen?

$$k = \frac{\text{Rel*} \cdot (1\text{-Rel})}{\text{Rel} \cdot (1 - \text{Rel*})} = \frac{0.95 \cdot (1 - 0.88)}{0.88 \cdot (1 - 0.95)} = 2.59$$

3.3 Streuung der Testwerte

- Eine hohe Streuung geht meist mit einer hohen Reliabilität einher, während bei geringer Streuung eine hohe Reliabilität unwahrscheinlich ist
- Beispiel: Erfassung von Intelligenz in der Allgemeinbevölkerung vs. in der Subpopulation der Psychologiestudierenden
- Populationsabhängigkeit der Reliabilität weist auf generelles Problem der Konzeption der Reliabilität in der KTT hin: Es gibt nur *eine* Messgenauigkeit, die pauschal für alle potenziellen Testwerte gilt
- In der IRT kann dagegen die Messgenauigkeit einzelner Testwerte angegeben werden (→ Testinformation)

Reliabilität

1. Definition Reliabilität
2. Methoden zur Reliabilitätsschätzung
3. Einflussfaktoren auf die Höhe der Reliabilität
4. **Anwendung: Konfidenzintervalle in der Individualdiagnostik**

4. Konfidenzintervalle

- Der beobachtete Testwert (z.B. Summenscore) ist ein Punktschätzer für den wahren Wert
- In der Individualdiagnostik ist es wichtig zu wissen, wie präzise die Schätzung ist
- Der Standardmessfehler ist ein Maß für die Präzision der Messung: Er gibt an, wie stark die Messfehler um die wahren Werte streuen
- $$s_e = s_x \cdot \sqrt{1 - \mathrm{Rel}(x)}$$

  s_x: Standardabweichung der beobachteten Werte

4. Konfidenzintervalle

- Beispiel Narzissmus-Test:

$$s_e = 14.16 \cdot \sqrt{1 - 0.91} = 4.25$$

- In der Regel werden normierte Werte verwendet, z. B. T-Werte mit *M* = 50 und *SD* = 10

$$s_e = 10 \cdot \sqrt{1 - 0.91} = 3$$

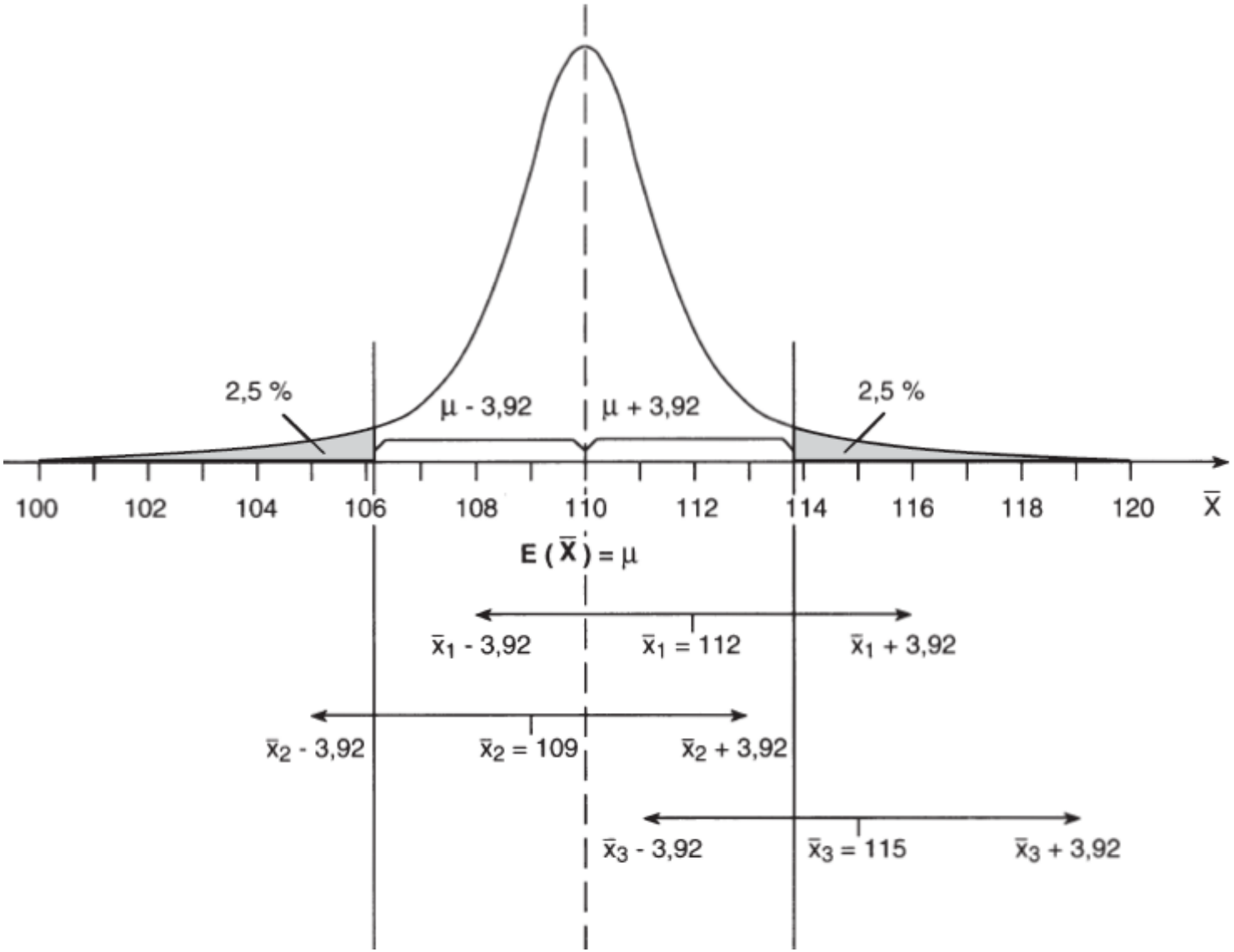
4. Konfidenzintervalle

- Mit s_e kann ein Wertebereich berechnet werden, der den wahren Wert mit einer gewissen Wahrscheinlichkeit überdeckt: Das Konfidenzintervall (KI)

$$\mathrm{KI} = X \pm z_{\alpha/2} \cdot s_e$$

- Der Konfidenzkoeffizient von 1 – α (z.B. 95 oder 99) ist die Wahrscheinlichkeit, mit der die Schätzung zu Intervallen führt, die den wahren Wert enthalten
- Der z-Wert $z_{\alpha/2}$ ist der Wert, der von der Standard-normalverteilung α/2 abschneidet
- Z. B. liegen 95% der Fläche unter der Normalverteilung zwischen den z-Werten -1.96 und +1.96

4. Konfidenzintervalle



4. Konfidenzintervalle

- Beispiel Narzissmus-Test:
  - Eine Person hat einen T-Wert von 60
  - Wir wählen einen Konfidenzkoeffizient von 95

$$\mathrm{KI} = X \pm z_{\alpha/2} \cdot s_e = 60 \pm 1.96 \cdot 3 = 60 \pm 5.88$$

  - Untere und obere Grenze des Konfidenzintervalls:

$$\mathrm{KI_u} = 60 - 5.88 = 54.12$$

$$\mathrm{KI_o} = 60 + 5.88 = 65.88$$

  - Interpretation: Mit 95%iger Wahrscheinlichkeit gehört das berechnete Konfidenzintervall zu denjenigen Intervallen, die den wahren Wert enthalten → Es ist sehr plausibel, dass der wahre Wert der Person zwischen 54.12 und 65.88 liegt

4. Konfidenzintervalle

- Konfidenzintervalle sind wichtig, um Aussagen über die Ausprägung einer Person zu treffen, die die Präzision (bzw. Unreliabilität) der Messung mitberücksichtigen
- Z.B. könnte man den Durchschnittsbereich als M +/- 1SD definieren (für T-Werte 40-60)
- Eine Person mit einem T-Wert von 61 würde man – alleine basierend auf dem Punktschätzer – als überdurchschnittlich narzisstisch beschreiben
- Berücksichtigt man dagegen den s_e, wären auch durchschnittliche Werte für die Person plausibel:

$$\mathrm{KI} = X \pm z_{\alpha/2} \cdot s_e = 61 \pm 1.96 \cdot 3 = 61 \pm 5.88$$

- Man würde sie daher als durchschnittlich bis überdurchschnittlich narzisstisch beschreiben

4. Konfidenzintervalle

- In der KTT wird für jeden Testwert der gleiche Standardmessfehler verwendet
- In der IRT gibt es verschiedene Standardmessfehler in Abhängigkeit von der Traitausprägung θ, die mithilfe der Testinformation ermittelt werden können

$$s_e(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

- Damit kann die Breite der Konfidenzintervalle für Personen mit unterschiedlichen Traitausprägungen variieren

4. Konfidenzintervalle

Fall	Summen-score	Max. Summen-score	Personen-parameter	SE Personen-parameter
1	8.00	9.00	2.20986	1.02769
2	8.00	9.00	2.20986	1.02769
3	9.00	9.00	3.54575	1.64382
4	6.00	9.00	0.80090	0.79482
5	8.00	9.00	2.20986	1.02769
6	6.00	9.00	0.80090	0.79482
7	5.00	9.00	0.23281	0.76253
8	7.00	9.00	1.43297	0.86334
9	6.00	9.00	0.80090	0.79482
10	5.00	9.00	0.23281	0.76253
11	7.00	9.00	1.43297	0.86334
12	6.00	9.00	0.80090	0.79482
13	7.00	9.00	1.43297	0.86334
14	6.00	9.00	0.80090	0.79482
15	8.00	9.00	2.20986	1.02769
16	3.00	9.00	-0.84698	0.77567
17	9.00	9.00	3.54575	1.64382

KI für θ = 2.2:

$$KI = 2.2 \pm 1.96 \cdot 1.03 =$$

$$2.2 \pm 2.02$$

KI für θ = 0.8:

$$KI = 0.8 \pm 1.96 \cdot 0.79 =$$

$$0.8 \pm 1.55$$

Literatur zu dieser Sitzung

Moosbrugger & Kelava (2012). Kapitel 6.

Moosbrugger & Kelava (2012). Kapitel 5.5.2 und 5.6.