

Testtheorie und Testkonstruktion

WS 2020/21

9. Objektivität

18.01.2021

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

Objektivität

Ein Test ist dann objektiv, wenn die Durchführung und Auswertung des Tests sowie die Interpretation des Testergebnisses unabhängig von der Testleiterin/dem Testleiter ist.

1. Durchführungsobjektivität
2. Auswertungsobjektivität
3. Interpretationsobjektivität

2. Auswertungsobjektivität

Maße zur Beurteilung der Übereinstimmung zwischen Beobachter*innen (Interrater-Reliabilität)

2.1 Für nominalskalierte Daten:

Cohens Kappa (1960)

2.2 Für intervallskalierte Daten:

Intraklassen-Korrelation (intraclass correlation; ICC)

2.1 Cohens Kappa

Beispiel: Zwei Beobachter beurteilen Führungskräfte hinsichtlich ihres Führungsstils

Beurteilungskategorien: Charismatisch, transformational, transaktional

Anteil der Fälle, die von Beobachter A und B den 3 Kategorien zugeordnet werden:

		Beobachterin A			
	Kategorie	1	2	3	
Beobachter B	1	.44	.07	.09	.60
	2	.05	.20	.05	.30
	3	.01	.03	.06	.10
		.50	.30	.20	$\Sigma 1.00$

2.1 Cohens Kappa

		Beobachterin A			
	Kategorie	1	2	3	
Beobachter B	1	.44	.07	.09	.60
	2	.05	.20	.05	.30
	3	.01	.03	.06	.10
		.50	.30	.20	$\Sigma 1.00$

Einfachster Ansatz zur Bestimmung der Beobachterübereinstimmung:

$$p_o = .44 + .20 + .06 = .70$$

Warum ist dieser Ansatz inadäquat?

2.1 Cohens Kappa

		Beobachterin A			
	Kate- gorie	1	2	3	
Beo- bachter B	1	.44 (.30)	.07	.09	.60
	2	.05	.20 (.09)	.05	.30
	3	.01	.03	.06 (.02)	.10
		.50	.30	.20	$\Sigma 1.00$

$$p_e = .30 + .09 + .02 = .41$$

2.1 Cohens Kappa

- Cohens Kappa berücksichtigt nur den Anteil der beobachteten Übereinstimmung, der über die per Zufall erwartete Übereinstimmung hinausgeht:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

- Damit repräsentiert Cohens Kappa die um zufällige Übereinstimmungen korrigierte Beobachterübereinstimmung
- Berechnung mit Häufigkeiten:

$$\kappa = \frac{f_o - f_e}{N - f_e}$$

2.1 Cohens Kappa

- Beispiel:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{.70 - .41}{1 - .41} = .49$$

- Voraussetzungen:
 - Untersuchungseinheiten sind unabhängig
 - Kategorien sind disjunkt und exhaustiv
 - Beobachter*innen geben ihre Urteile unabhängig voneinander ab

2.1 Cohens Kappa

- **Wertebereich von Kappa:**

- Stimmt p_o mit p_e überein, ist $\kappa = 0$
- Bei perfekter Übereinstimmung (Elemente neben Diagonale = 0) zwischen den Beobachter*innen ist $\kappa = 1$
- Ungleiche Randverteilung und/oder Elemente neben Diagonale $\neq 0$: Maximalwert für κ ist < 1

$$\kappa_{\max} = \frac{p_{o,r} - p_e}{1 - p_e}$$

$p_{o,r}$: Summe der kleineren Randanteile

2.1 Cohens Kappa

		Beobachterin A			
	Kategorie	1	2	3	
Beobachter B	1	.50 (.25)	.0	.0	.50
	2	.0	.30 (.09)	.0	.30
	3	.0	.0	.20 (.04)	.20
		.50	.30	.20	Σ1.00

$$\kappa_{\max} = \frac{p_{o,r} - p_e}{1 - p_e} = \frac{1 - .38}{1 - .38} = 1$$

		Beobachterin A			
	Kategorie	1	2	3	
Beobachter B	1	.44 (.30)	.07	.09	.60
	2	.05	.20 (.09)	.05	.30
	3	.01	.03	.06 (.02)	.10
		.50	.30	.20	Σ1.00

$$\begin{aligned} \kappa_{\max} &= \frac{p_{o,r} - p_e}{1 - p_e} \\ &= \frac{(.50 + .30 + .10) - .41}{1 - .41} \\ &= \frac{.90 - .41}{1 - .41} = .83 \end{aligned}$$

2.1 Cohens Kappa

- Beurteilung von Kappa:
 - $\leq .40$: ungenügend
 - $.41 - .60$: befriedigend
 - $.61 - .80$: gut
 - $>.80$: sehr gut

2.1 Cohens Kappa

Wenn nicht jeder Rater jede Person beurteilt, sollten die fehlenden Ratings von der Berechnung der beobachteten Übereinstimmung entfernt werden

		Beobachterin A				
	Kategorie	1	2	3	X	
Beobachter B	1	.18 (.09)	.08	.02	.03	.31
	2	.05	.23 (.14)	.09	.02	.38
	3	.02	.04	.14 (.07)	.03	.23
	X	.03	.01	.05	0 (.006)	.08
		.28	.36	.30	.07	Σ1.00

$$p_o = \frac{.18 + .23 + .14}{1 - (.07 + .08)} = 0.65$$

Die weitere Berechnung von Cohens Kappa erfolgt wie gewohnt.

2.1 Cohens Kappa

Cohens Kappa kann in 2 Situationen zu paradoxen Werten führen:

1. Bei sehr hohen erwarteten Anteilen führt die Korrektur zu sehr niedrigen Kappa-Werten
2. Ungleiche, asymmetrische Randverteilungen können zu höheren Kappa-Werten führen als ähnliche, symmetrische Randverteilungen

2.1 Cohens Kappa

Paradoxon 1

Es werden 100 Schüler*innen auf Hochbegabung getestet

		Beobachterin A		
	Kategorie	Nicht hochbegabt	Hochbegabt	
Beobachter B	Nicht hochbegabt	.90	.10	1.00
	Hochbegabt	0	0	0
		.90	.10	$\Sigma 1.00$

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{.90 - .90}{1 - .90} = 0$$

2.1 Cohens Kappa

Paradoxon 2

Im Assessment Center werden 100 Bewerber*innen hinsichtlich ihrer Eignung beurteilt

		Beobachterin A		
	Kategorie	Nicht geeignet	Ge-eignet	
Beobachter B	Nicht geeignet	.45	.15	.60
	Ge-eignet	.25	.15	.40
		.70	.30	Σ1.00

		Beobachterin A		
	Kategorie	Nicht geeignet	Ge-eignet	
Beobachter B	Nicht geeignet	.25	.35	.60
	Ge-eignet	.5	.35	.40
		.30	.70	Σ1.00

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{.60 - .54}{1 - .54} = .13$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{.60 - .46}{1 - .46} = .26$$

2. Auswertungsobjektivität

Maße zur Beurteilung der Übereinstimmung zwischen Beobachter*innen (Interrater-Reliabilität)

2.1 Für nominalskalierte Daten:

Cohens Kappa (1960)

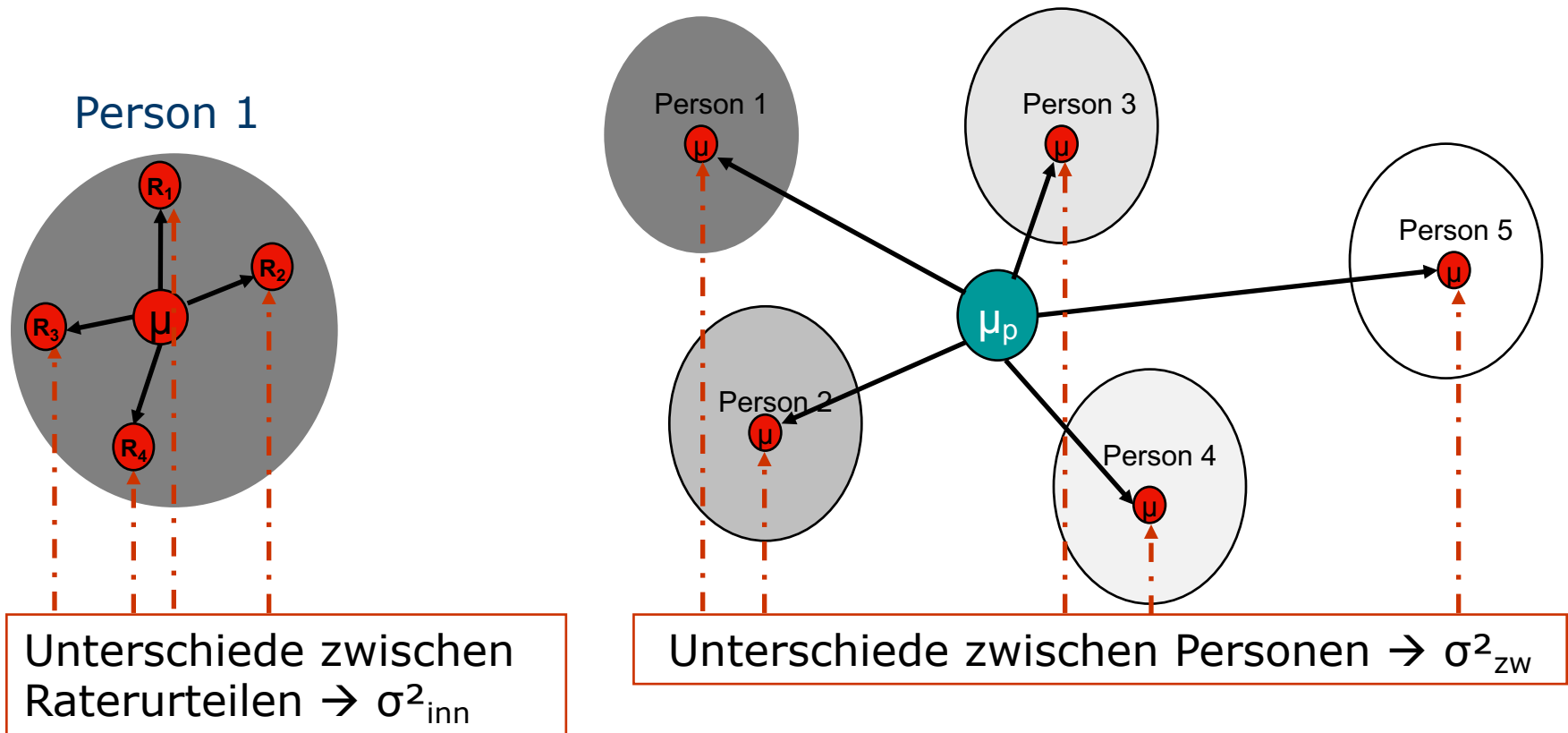
2.2 Für intervallskalierte Daten:

Intraklassen-Korrelation (intraclass correlation; ICC)

2.2 ICC

- Bei der Intraklassenkorrelation (ICC) wird die Beobachterübereinstimmung mithilfe von Varianzverhältnissen quantifiziert
- Hintergrund: Ist die Varianz zwischen Objekten groß im Vergleich zu der Varianz innerhalb von Objekten, spricht dies für eine hohe Beobachterübereinstimmung

2.2 ICC



2.2 ICC

- Es gibt 6 verschiedene ICC Koeffizienten
- In Abhängigkeit von der Datenkonstellation wird der angemessene Koeffizient ausgewählt
 - Zufallsauswahl der Beobachter – ja/nein?
 - Wird jedes Objekt von jedem Beobachter beurteilt – ja/nein?
 - Werden Ratings einzelner Beobachter oder ein Mittelwert aus den Ratings verschiedener Beobachter ausgewertet?

2.2 ICC

- Bei der Konstellation Ja – Nein – Mittelwert aus k Ratings wird ICC(1,k) verwendet:

$$ICC(1,k) = \frac{\hat{\sigma}_{zw}^2 - \hat{\sigma}_{inn}^2}{\hat{\sigma}_{zw}^2}$$

$\hat{\sigma}_{zw}$ = Varianz zwischen den Objekten

$\hat{\sigma}_{inn}$ = Varianz innerhalb der Objekte

2.2 ICC

- Bei der Konstellation Ja – Nein – einzelne Ratings wird ICC(1,1) verwendet:

$$ICC(1,1) = \frac{\hat{\sigma}_{zw}^2 - \hat{\sigma}_{inn}^2}{\hat{\sigma}_{zw}^2 + (k-1)\hat{\sigma}_{inn}^2}$$

k = Anzahl der Rater

$\hat{\sigma}_{zw}$ = Varianz zwischen den Objekten

$\hat{\sigma}_{inn}$ = Varianz innerhalb der Objekte

2.2 ICC

Beispiel: 6 Bewerber*innen werden während einer Gruppendiskussion in einem Assessment Center von jeweils 4 von 8 zufällig ausgewählten Ratern hinsichtlich ihrer Extraversion auf einer Skala von 1 – 10 eingeschätzt

		Rating			
		A	B	C	D
Bewerber/in	1	5	6	5	3
	2	8	9	4	6
	3	7	5	5	8
	4	2	3	3	1
	5	8	9	9	7
	6	1	3	2	4

2.2 ICC

		Rating				
		A	B	C	D	M
Bewer- ber/in	1	5	6	5	3	4.75
	2	8	9	4	6	6.75
	3	7	5	5	8	6.25
	4	2	3	3	1	2.25
	5	8	9	9	7	8.25
	6	1	3	2	4	2.50
						5.125

$\hat{\sigma}_{zw}$: Wie stark streuen die M der Bewerber*innen um den Gesamtmittelwert?

$\hat{\sigma}_{inn}$: Wie stark streuen die Ratings um den Personenmittelwert?

2.2 ICC

Verwendung von Formeln der einfaktoriellen Varianzanalyse zur Berechnung der Varianz zwischen den Bewerber*innen und der Varianz innerhalb der Bewerber*innen:

$$QS_{zw} = \sum_{j=1}^J n_j \cdot (\bar{x}_j - \bar{x})^2 \quad df_{zw} = J - 1$$

$$\begin{aligned} QS_{zw} &= 4 \cdot [(4.75 - 5.125)^2 + (6.75 - 5.125)^2 + (6.25 - 5.125)^2 + \\ &\quad (2.25 - 5.125)^2 + (8.25 - 5.125)^2 + (2.50 - 5.125)^2] \\ &= 115.875 \end{aligned}$$

$$\hat{\sigma}_{zw}^2 = \frac{QS_{zw}}{J-1} = \frac{115.875}{5} = 23.175$$

2.2 ICC

$$QS_{inn} = \sum_{j=1}^J \sum_{m=1}^{n_j} (x_{mj} - \bar{x}_j)^2$$

$$df_{inn} = n - J$$

$$\begin{aligned} QS_{inn} &= (5 - 4.75)^2 + (6 - 4.75)^2 + (5 - 4.75)^2 + \\ &(3 - 4.75)^2 + \dots + (4 - 2.50)^2 \\ &= 36.75 \end{aligned}$$

$$\hat{\sigma}_{inn}^2 = \frac{QS_{inn}}{n - J} = \frac{36.75}{18} = 2.04$$

$$ICC(1,k) = \frac{\hat{\sigma}_{zw}^2 - \hat{\sigma}_{inn}^2}{\hat{\sigma}_{zw}^2} = \frac{23.175 - 2.04}{23.175} = 0.91$$

Literatur zu dieser Sitzung

Gwet (2010): Handbook of inter-rater reliability. Kapitel 2
(prüfungsrelevant: S. 11-25 und 30-34)