Grundlagen der Diagnostik

Sitzung 5

Beobachter- und Beurteilerübereinstimmung II

Lehrstuhl für Psychologische Methodenlehre und Diagnostik der Ludwig-Maximilians-Universität München

Vorläufiger Semesterüberblick

UPDATE

Vorlesung Grundlagen der Diagnostik SoSe 26

Sitzung	Datum	Thema	Themenblock
1	16.04.26	Einführung I	Definitionen; diagnostischer Prozess; gesetzlicher Rahmen; diagnostische Entscheidungen; Gütekriterien
2	23.04.26	Einführung II	
3	30.04.26	Einführung II fertig	
4	07.05.26	Verhaltensbeobachtung	Verhaltensbeobachtung als diagnostisches Verfahren
/	14.05.26	entfällt wegen Feiertag	
5	21.05.26	Beobachterübereinstimmung I	Maße zur Bestimmung der Übereinstimmung zwischen verschiedenen Beobachtern/Ratern
6	28.05.26	Beobachterübereinstimmung II	
/	04.06.26	*entfällt wegen Feiertag*	
7	11.06.26	Interviews	Interviews als diagnostisches Verfahren
8	18.06.26	Urteile und Fehler	Diagnostische Urteilsbildung und Güte von Urteilen
9	25.06.26	Einzelfalldiagnostik I	Methoden der Einzelfalldiagnostik aus der frequentistischen und bayesianischen Statistik
10	02.07.26	Einzelfalldiagnostik II	
11	09.07.26	Digitale Diagnostik	Diagn. Einsatz von digitalen Daten und maschinellem Lernen
12	16.07.26	Fragestunde (via Zoom)	Ihre Fragen zur Vorlesung und dem UK

Vorläufiger Semesterüberblick

Sitzung	Datum	Thema	Themenblock
1	16.04.26	Einführung I	Definitionen; diagnostischer Prozess; gesetzlicher Rahmen; diagnostische Entscheidungen; Gütekriterien
2	23.04.26	Einführung II	
3	30.04.26	Einführung II fertig	
4	07.05.26	Verhaltensbeobachtung	Verhaltensbeobachtung als diagnostisches Verfahren
/	14.05.26	entfällt wegen Feiertag	
5	21.05.26	Beobachterübereinstimmung I	Maße zur Bestimmung der Übereinstimmung zwischen verschiedenen Beobachtern/Ratern
6	28.05.26	Beobachterübereinstimmung II	

➛ In der heutigen Vorlesung befassen wir uns wieder mit Maßen der Beobachterübereinstimmung – diesmal für ordinale Ratings

Verschiedene Indizes für Beobachterübereinstimmung, abhängig vom Skalenniveau…

Bei nominalskalierten Daten:

- Prozentuale Übereinstimmung
- Cohens κ (kappa) und Scotts π (pi)
- Odds Ratio (auch: „Risikoverhältnis“) und Yules Y

Bei ordinal- und intervallskalierten Daten:

- Rangkorrelationen
- Intra-Klassen-Korrelation (Intra-Class-Correlation, ICC)

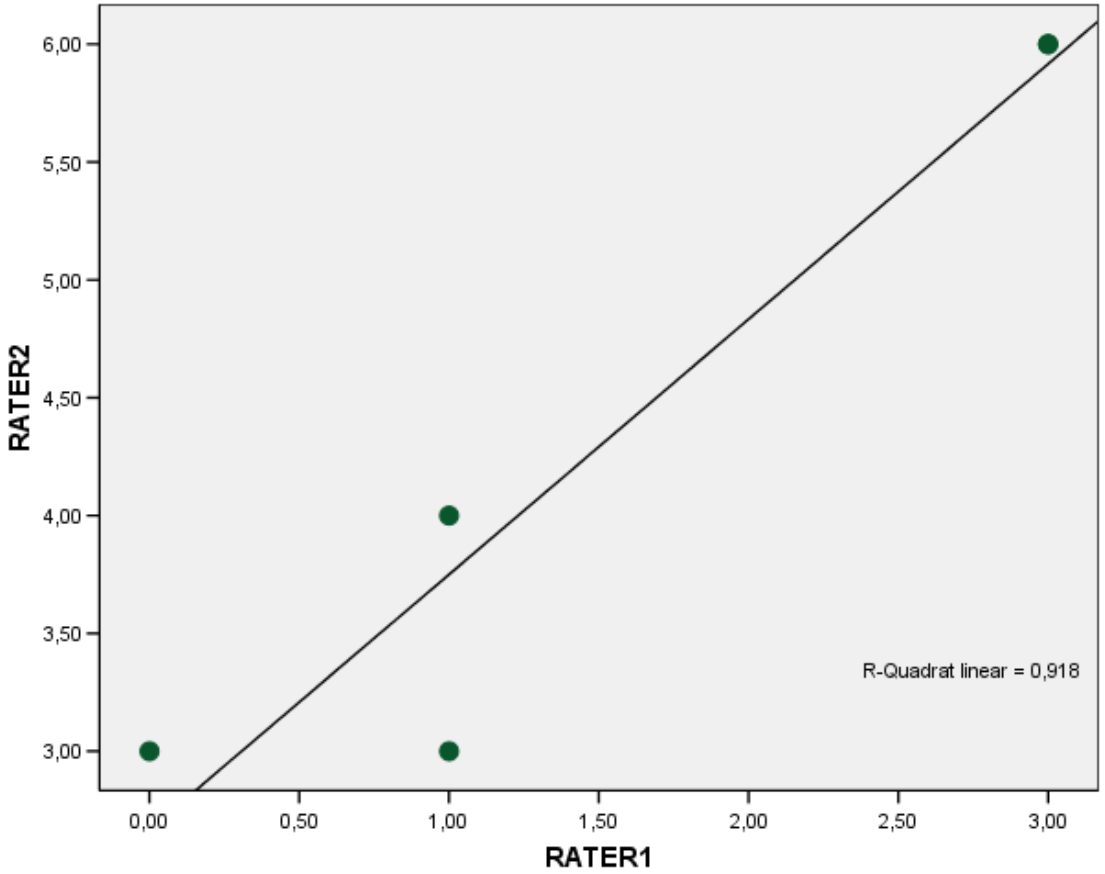
4. Überblick über Maße für ordinal- und intervallskalierte Daten

Möglichkeit 1: „Normale“ Korrelationskoeffizienten

- Die Korrelation zwischen beiden Ratern als Schätzung für die Übereinstimmung verwenden
- Die Pearson-Korrelation zwischen beiden Ratern beträgt r = .96

Aber:

- Nur 2 Rater
- Justiert (nur Rangreihe)



Rückblick: Unjustierte und Justierte Maße

Unjustierte Maße:

- Bestrafen unterschiedliche **Strenge** (Wahrnehmungsschwellen bei nominalskalierten Daten)
- Bestrafen mangelnde **Konsistenz**
- Konsequenz: **Absolute Übereinstimmung** wird bewertet

Justierte Maße:

- Bestrafen ***nicht*** unterschiedliche **Strenge**
- Bestrafen ***ausschließlich*** mangelnde **Konsistenz**
- Konsequenz: Nur Einhaltung der Rangreihe wird bewertet („**relative Übereinstimmung**“)

Konsistenz vs. Strenge bei Ordinal-/Intervallskalen

Beobachtung	Beurteiler A	Beurteiler B
#1	1	3
#2	1	3
#3	2	4
#4	2	4
#5	3	5
#6	3	5

→ konsistent, aber nicht gleich streng

Möglichkeit 2: Intra-Klassen-Korrelationen

Varianzanalytischer Ansatz:

- Erklärung der Unterschiede zwischen den realisierten Messwerten (z.B. Beurteilungen): $VAR(X_i)$
- Wiederholung:
  - Klassische Testtheorie (→ VL2 Testtheorie): $VAR(X_i) = VAR(\tau_i) + VAR(\varepsilon_i)$
  - Reliabilität (→ VL10 Testtheorie): $REL(X_r) = \frac{VAR(\tau_i)}{VAR(X_i)} = \frac{VAR(\tau_i)}{VAR(\tau_i)+VAR(\varepsilon_i)}$
- ICCs sind eine Schätzung der Reliabilität basierend auf (verschiedenen) Schätzungen für $VAR(\tau_i)$ und $VAR(\varepsilon_i)$
- Uneinigkeiten der Rater bei der Beurteilung werden in $VAR(\varepsilon_i)$ erfasst
- Die Varianz der „wahren Werte“ $VAR(\tau_i)$ wurde in Testtheorie auch als die „systematischen“ Unterschiede bezeichnet. Wir erinnern uns, diese ist nicht immer identisch ist mit den wahren Unterschieden der Personen im Merkmal: $VAR(\theta)$

Möglichkeit 2: Intra-Klassen-Korrelationen

Varianzanalytischer Ansatz (Achtung: i steht jetzt für Rater!):

- Zusätzliche Möglichkeit der Berücksichtigung der Ratervarianz (→ zweifaktorielle Varianzanalyse):

$$VAR(X_i) = VAR(\theta) + VAR(rater_i) + VAR(\varepsilon_i)$$

- $VAR(\theta)$ ist die Varianz der wahren Unterschiede zwischen den Personen
- $VAR(rater_i)$ beschreibt systematische Unterschiede zwischen den Ratern
  - Werden Raterunterschiede modelliert, ist $VAR(rater_i)$ ein Faktor in der Analyse, welcher Varianz in den Messwerten (also in $VAR(X_i)$), erklären kann (= eine weitere UV)
  - Werden die Raterunterschiede nicht separat modelliert, ist $VAR(rater_i)$ in $VAR(\varepsilon_i)$ enthalten (systematische Raterunterschiede als Teil der „Fehler“)

Hinweis: Eine Interaktion zwischen Personen und Ratern wäre theoretisch denkbar, und damit auch eine Varianz $VAR(\theta*\text{rater})$ dieser Interaktion, aber diese wird typischerweise außen vor gelassen (d.h. als 0 betrachtet)

Möglichkeit 2: Intra-Klassen-Korrelationen

Varianzanalytischer Ansatz:

- Zusätzliche Möglichkeit der Berücksichtigung der Ratervarianz (→ zweifaktorielle Varianzanalyse):

$$VAR(X_i) = VAR(\theta) + VAR(rater_i) + VAR(\varepsilon_i)$$

- Wichtig: Es sind getrennte Entscheidungen, ob man …
  - … die Ratervarianz **modelliert** (d.h. im Modell aufnimmt, und dann durch Schätzung quantifiziert)
  - … die Ratervarianz **als Fehler berücksichtigt** (vgl. später ICC-Modell 2 und 3)

Höhe von ICC-Koeffizienten

Theoretischer Wertebereich: -1 bis 1 (praktisch < 0 = 0)

- 0 = Varianz ist ausschließlich auf Messfehler zurückzuführen (keine Reliabilität)
- 1 = Varianz ist ausschließlich auf wahre Werte zurückzuführen (perfekte Reliabilität)

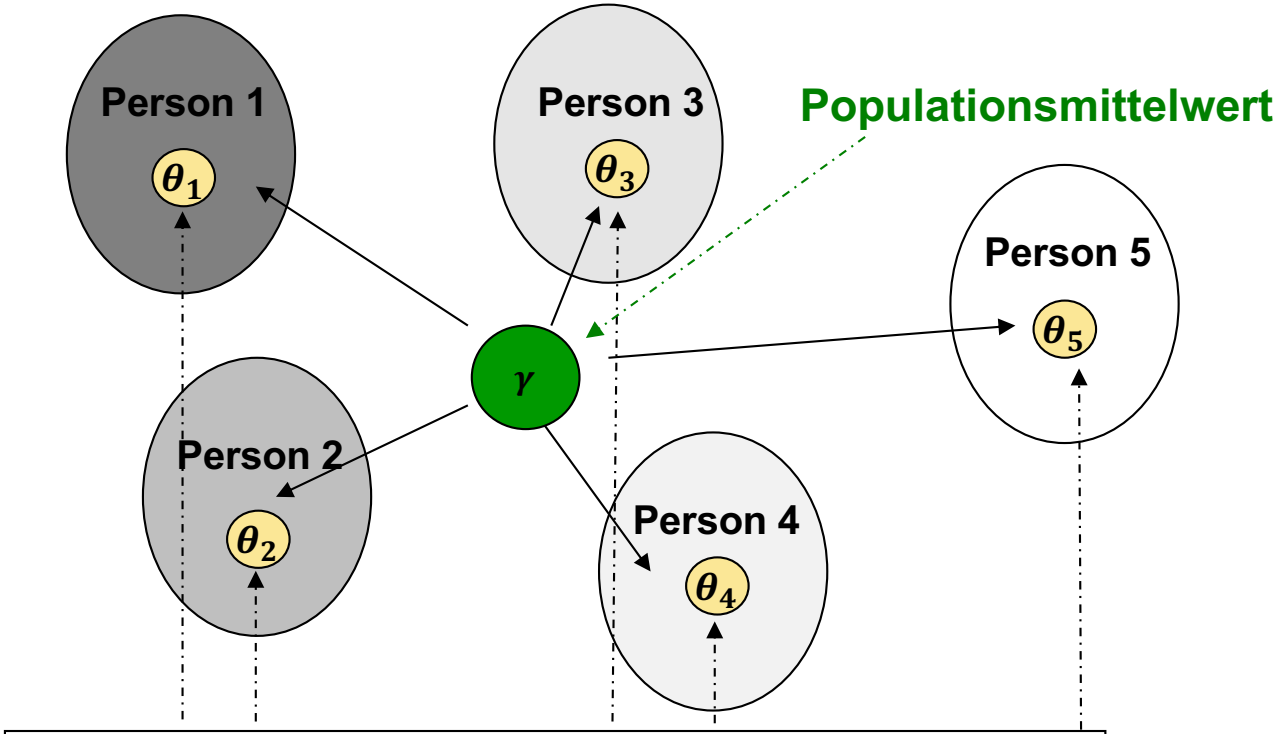
Faustregeln zur Beurteilung (Fleiss, 1981; Cicchetti & Sparrow, 1981):

< 0.40	=	schlecht
0.40 – 0.59	=	befriedigend
0.60 – 0.74	=	gut
> 0.74	=	sehr gut

Bedeutung der verschiedenen Varianzquellen

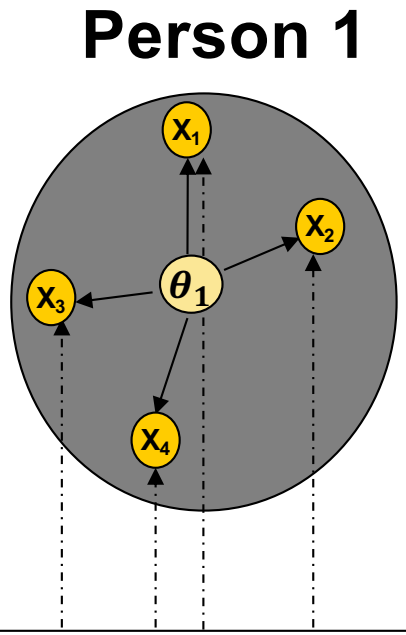
- **Faktor 1 = wahre Unterschiede**
  - Wahre Unterschiede zwischen den Personen = Schätzung der Varianz im Merkmal (bzw. der latenten Variable) → $VAR(\theta)$
- **Optionaler Faktor 2 = Raterunterschiede**
  - Unterschiede zwischen den mittleren Ratings der Rater über alle Personen hinweg = Schätzung unterschiedlicher Strenge (Wahrnehmungsschwellen) → $VAR(rater_i)$
- **Fehler**
  - Unterschiede der Rater in den Bewertungen für die Personen, die auf (sonstige) Fehler zurückgehen = Schätzung Fehlervarianz der Messwerte → $VAR(\varepsilon_i)$

Intra-Klassen-Korrelationen: Schätzung der verschiedenen Varianzquellen



Tatsächliche Unterschiede zwischen Personen
→ $VAR(\theta)$

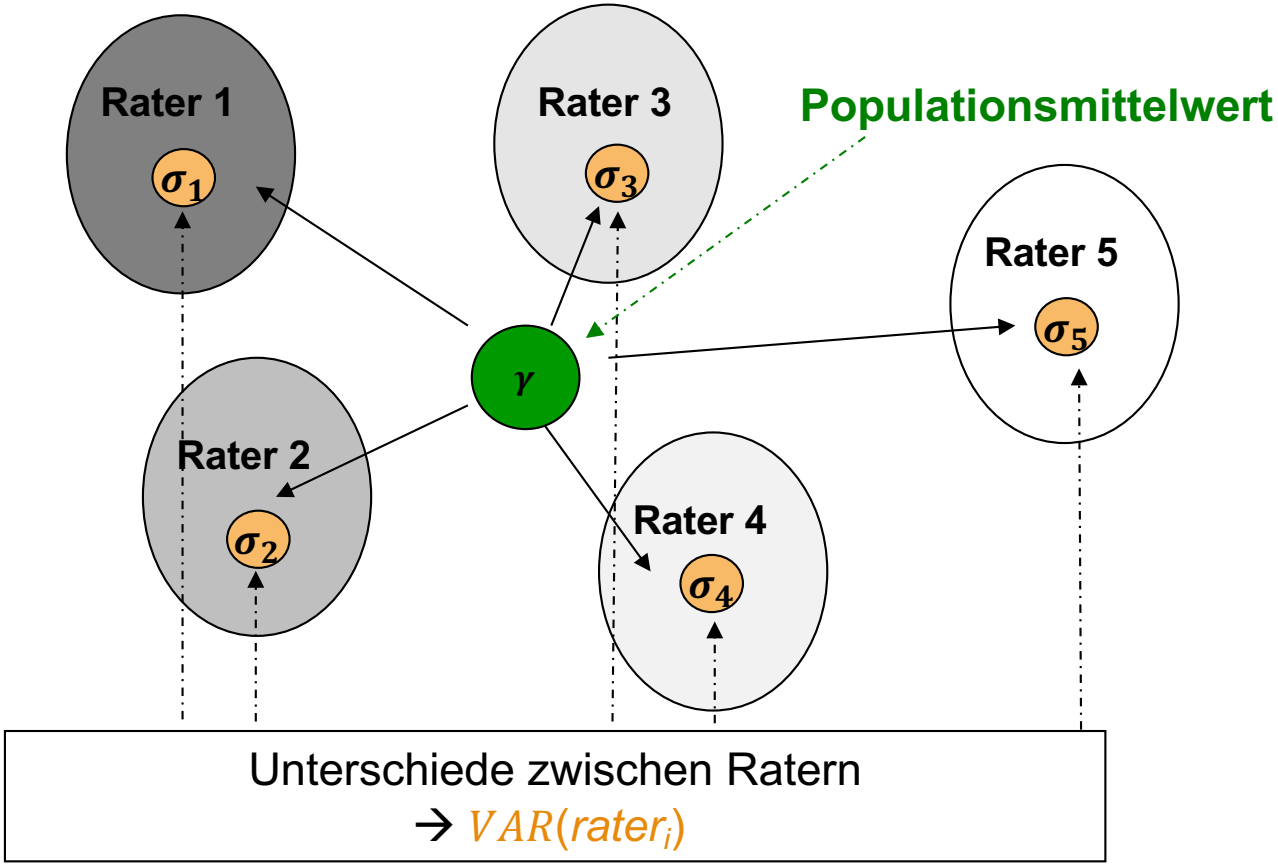
große Kreise = „Raterraum“
= Mögliche Raterurteile



Abweichung der Ratings von den jeweiligen wahren Werten
→ $VAR(\varepsilon_i)$

- Ist $VAR(\theta)$ groß, dann unterscheiden sich die Personen bezüglich des Merkmals stark
- Ist $VAR(\varepsilon_i)$ groß, dann unterscheiden sich einzelne Ratings für die gleiche Person stark

Intra-Klassen-Korrelationen: Schätzung der verschiedenen Varianzquellen



- Ist $VAR(rater_i)$ groß, dann unterscheiden sich die Rater in ihrer mittleren Bewertung stark

- Die Varianz der wahren Unterschiede im Merkmal $VAR(\theta)$, Ratervarianz $VAR(rater_i)$ und Fehlervarianz $VAR(\varepsilon_i)$ sind nicht beobachtbar und müssen in einem statistischen Modell geschätzt werden.
- Die ursprüngliche Schätzmethode basiert auf varianzanalytischen Modellen (ein- und zweifaktorielle ANOVAs). Heute werden in der Regel sogenannte **Gemischte Lineare Modelle** (Linear Mixed Models, LMMs) verwendet.
- Mithilfe dieser Modelle erhalten wir die Schätzwerte: $\widehat{VAR(\theta)}$, $\widehat{VAR(rater_i)}$ und $\widehat{VAR(\varepsilon_i)}$

5. Eigenschaften unterschiedlicher ICCs

Eigenschaften unterschiedlicher ICCs

Es gibt verschiedene ICC-Koeffizienten, mit unterschiedlicher Aussage und Interpretation (z.B. bzgl. Generalisierung, Art von Übereinstimmung)!

Zwei Szenarien sind hier relevant:

- Wenn man die Situation der Datenerhebung **noch gestalten** kann:
  - Überlegungen zu welcher Aussage man kommen möchte und Situation (z.B. Verhaltensbeobachtung) entsprechend planen
- Wenn die Situation / die Daten **schon gegeben** sind:
  - Eingeschränkte Entscheidung, welche ICCs möglich sind und welche Aussagekraft diese ICCs haben

Eigenschaften unterschiedlicher ICCs

→ Fragenkatalog für die Auswahl einer geeigneten ICC:

- 1-faktorielle vs. 2-faktorielle ICC
- Random vs. Fixed ICC
- Unjustierte vs. Justierte ICC
- Single vs. Average ICC

1-faktorielle vs. 2-faktorielle ICC

Wird die Varianz der Rater explizit modelliert oder nicht?

- Werden die Rater und deren Varianz $VAR(rater_i)$ als zweiter Faktor in das Modell aufgenommen?
  - Ja: ICC 2-faktoriell; Modell: $VAR(X_i) = VAR(\theta) + VAR(rater_i) + VAR(\varepsilon_i)$
  - Nein: ICC 1-faktoriell; Modell: $VAR(X_i) = VAR(\tau_i) + VAR(\varepsilon_i)$
- Achtung: Eine 2-faktorielle ICC kann nur berechnet werden, wenn es eine Überschneidung zwischen Personen und Ratern gibt. Im Idealfall wird jede Person von jedem Rater beurteilt (siehe Exkurs 1 im Anhang).

Hinweis: $VAR(\varepsilon_i)$ ist in den beiden Modellen numerisch nicht gleich, man könnte hier auch einen zusätzlichen „Varianznamen“ vergeben. $VAR(\varepsilon_i)$ im einfaktoriellen Modell entspricht $VAR(rater_i)$ + $VAR(\varepsilon_i)$ im zweifaktoriellen Modell.

Random vs. Fixed ICC

Möchte man die Ergebnisse hinsichtlich der Ratervarianz auf eine Population von Ratern generalisieren oder sind die untersuchten Rater die Einzigen von Interesse?

- ICC random: Die untersuchten Rater sind eine „repräsentative“ Stichprobe aus einer Population von Ratern → Generalisierung möglich
- ICC fixed: Die untersuchten Rater sind die einzigen, die es gibt → Aussage nur für diese Rater möglich
- Da sich diese Unterscheidung auf die Ratervarianz bezieht, unterscheiden sich nur 2-faktorielle ICCs in der Eigenschaft random / fixed!

Hinweis: Bei 1-faktoriellen ICCs fällt manchmal der Begriff „random“ - dies bezieht sich dann allerdings nicht auf die Ratervarianz (da diese nicht geschätzt wird; auch wenn man hier trotzdem von zufällig gezogenen Ratern ausgeht), sondern auf die bewerteten Personen, die in allen Modellen als „random“ (d.h. als Zufallsstichprobe) behandelt werden

Unjustierte vs. Justierte ICC

Interessiert man sich für die absolute Übereinstimmung oder nur die Rangreihe der Urteile?

- ICC unjustiert: absolute Übereinstimmung wird betrachtet
  → Beobachterinnen müssen konsistent (gleiche Rangreihe) *und* gleich streng urteilen (Übereinstimmung der absoluten Bewertung)
- ICC justiert: nur relative Übereinstimmung wird betrachtet
  → Es reicht, wenn Beobachterinnen die Personen in die gleiche Rangreihe bringen (d.h. konsistent sind)
- Dieser Unterschied kann als *Entscheidung* nur bei 2-faktoriellen ICCs einfließen. Eine 1-faktorielle ICC ist immer unjustiert.

Single vs. Average ICC

Interessiert man sich für die Reliabilität eines Urteils von nur einem Rater oder für die Reliabilität eines gemittelten Urteils von mehreren Ratern?

- ICC single: Wie gut ist die Übereinstimmung eines Urteils mit dem von anderen Ratern?
- ICC average: Wie messgenau ist das gemittelte Urteil mehrerer Rater?
- Dies hängt davon ab, ob Entscheidungen im jeweiligen Anwendungsfall basierend auf Einzelurteilen oder gemittelten Urteilen stattfinden
- Für alle ICCs sind die Betrachtungen von „single“ und „average“ möglich

Analogie zu Testtheorie: *Rater entsprechen Items*

- Single: Das Merkmal der Person wird nur mithilfe eines Raters (*eines Items*) gemessen.
- Average: Der Mittelwert von mehreren Ratern (*Itemmittelwert*) dient als Messwert für die Merkmalsausprägung der Person.

6. ICC Modelle nach Shrout & Fleiss (1979)

ICC Modelle (Shrout & Fleiss, 1979)

- Es gibt 3 verschiedene ICC-Modelle, die als Reliabilitätsschätzung herangezogen werden können
- Diese können jeweils als *single* und *average* ICC geschätzt werden, so dass sich nach Shrout & Fleiss (1979) insgesamt 6 ICC-Varianten mit unterschiedlichen Eigenschaften ergeben:

	ICC single	**ICC average**
ICC Modell 1 (1-faktoriell, unjustiert)	ICC(1,1)	ICC(1,k)
ICC Modell 2 (2-faktoriell, unjustiert, random)	ICC(2,1)	ICC(2,k)
ICC Modell 3 (2-faktoriell, justiert, fixed)	ICC(3,1)	ICC(3,k)

k = Anzahl der Ratings aus denen der Mittelwert gebildet wird,
z.B. bei 3 Ratern und Modell 1: ICC(1,3)

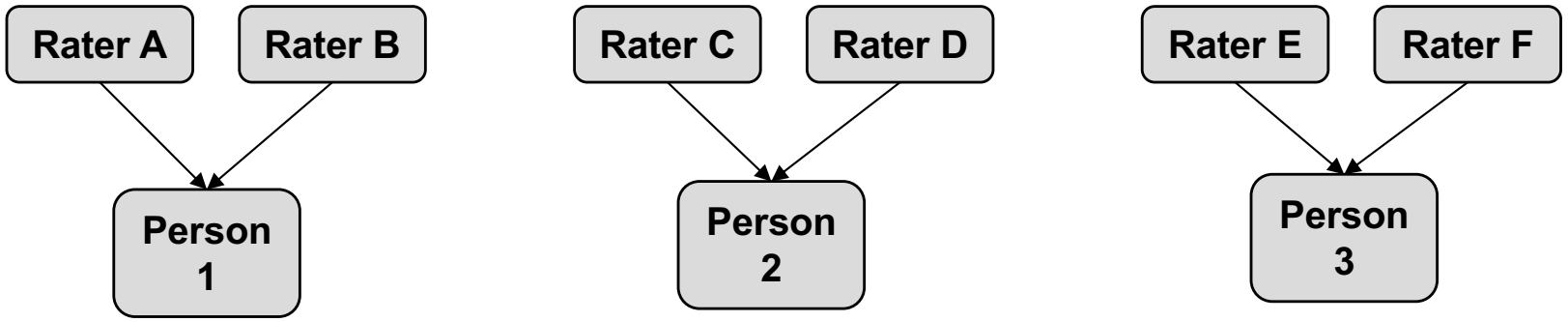
Exkurs: Erweiterungen der ICC-Modelle

- McGraw & Wong (1996) schlagen Varianten von den 2-faktoriellen Modellen 2 und 3 vor, bei denen unjustiert/justiert beliebig mit random/fixed kombiniert werden kann
- Diese sind allerdings nur auf Ebene der Interpretation als Varianten zu betrachten, nicht auf Ebene der mathematischen Modelle
- Nur die von Shrout & Fleiss aufgestellten Modelle mit den spezifischen Eigenschaftskombinationen (d.h. random für Modell 2 und fixed für Modell 3) können als Korrelationen im engeren Sinne interpretiert werden. Diese Interpretation als Korrelation schauen wir uns nur als Exkurs knapp an.
- Wir werden uns im Folgenden nur mit den Modellen von Shrout & Fleiss beschäftigen.

Modell „ICC1“: ICC 1-faktoriell & unjustiert

Anwendungs-Szenario:

- Es gibt einen Raterpool von *n* Ratern
- Jede Person wird von *unterschiedlichen* Raterkombinationen beobachtet, z.B.:

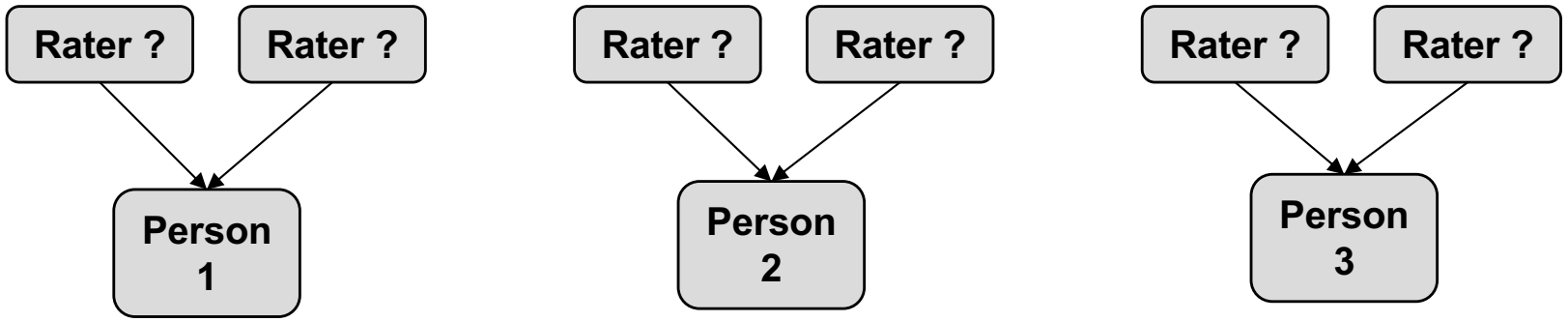


Hinweis: Das ICC1-Modell kann auf verschiedene Datensituationen angewendet werden, wird jedoch typischerweise für die obige Situation verwendet, da bei Vorliegen von Ratings aller Rater für alle Personen die anderen ICC-Modelle besser geeignet sind.

Modell „ICC1“: ICC 1-faktoriell & unjustiert

Alternatives Anwendungs-Szenario:

- Es gibt einen Raterpool von *n* Ratern
- Jede Person wird von *mehreren Ratern* beobachtet, aber man kann die Ratings *keinen Ratern zuordnen* (weil nicht dokumentiert) z.B.:



Eigenschaften:

- Unterschiede in der Varianz der Rater (d.h., Unterschiede in der Strenge) können nicht modelliert werden → 1-faktoriell, keine Unterscheidung von *fixed* / *random*
- Interpretation: Wie gut stimmen die Rater absolut überein? → unjustiert (Konsistenz & Strenge)

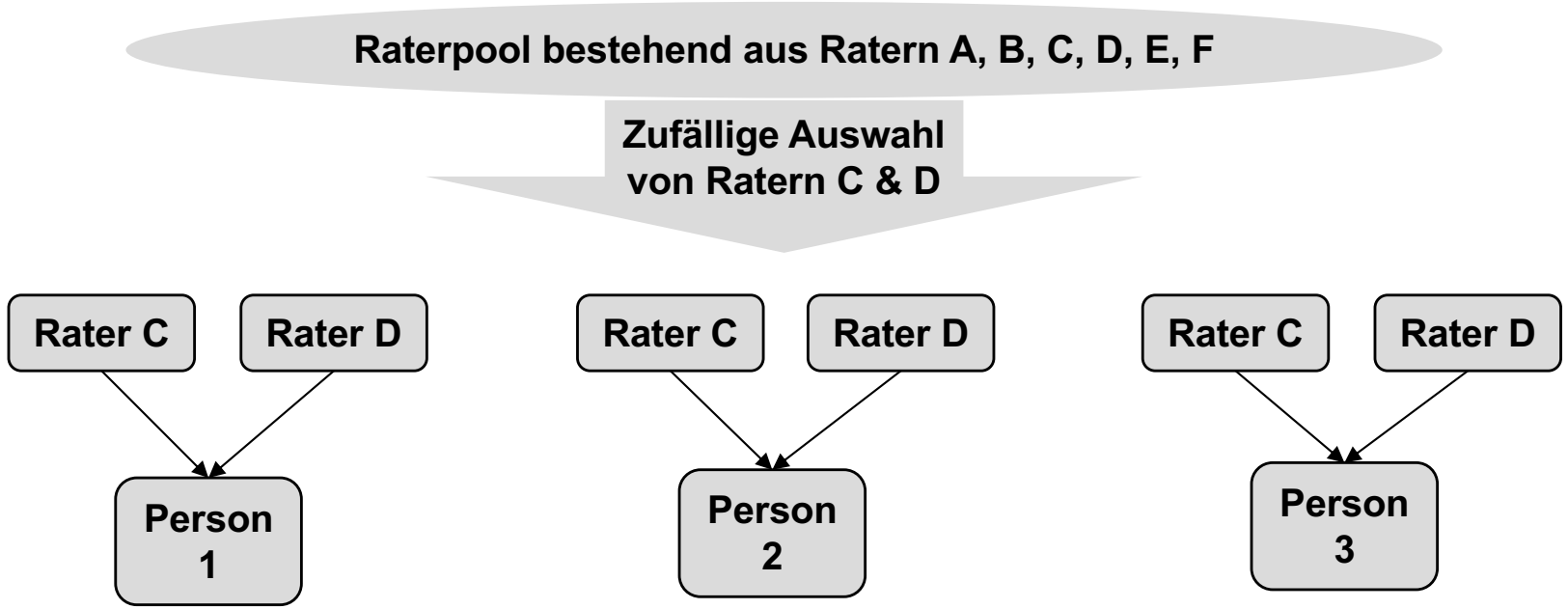
Varianten:

- Interessiert die Reliabilität eines Urteils → *single*
- Interessiert die Reliabilität des mittleren Urteils → *average*

Modell „ICC2“: ICC 2-faktoriell, random, unjustiert

Anwendungs-Szenario:

- Es gibt einen Raterpool von *n* Ratern
- Jede Person wird von der *gleichen* Raterkombination beobachtet, die zufällig aus dem Raterpool ausgewählt wurde (→ random), z.B.:



Prototypisches Beispiel:

- In einem Unternehmen werden Bewerberinnen in Assessment-Centern beurteilt
- Im Unternehmen gibt es 10 Leute, die eine Schulung für die Beurteilung in diesem Assessment-Center absolviert haben
- Innerhalb eines ACs wird jede Bewerberin von den gleichen, für dieses AC zufällig ausgewählten Ratern beurteilt
- Zwischen den ACs können die Rater wechseln, die berechneten Beurteilungsübereinstimmungsmaße sollen also generalisierbar auf alle möglichen Rater aus dem Raterpool sein

Eigenschaften:

- Varianz zwischen Ratern (d.h. Unterschiede in der Strenge) wird modelliert → 2-faktoriell
- Eine Generalisierung auf eine Population von Ratern ist vorgesehen → random
- Interpretation: Wie gut stimmen die Rater absolut überein? → unjustiert (Konsistenz & Strenge)

Varianten:

- Interessiert die Reliabilität eines Urteils → single
- Interessiert die Reliabilität des mittleren Urteils → average

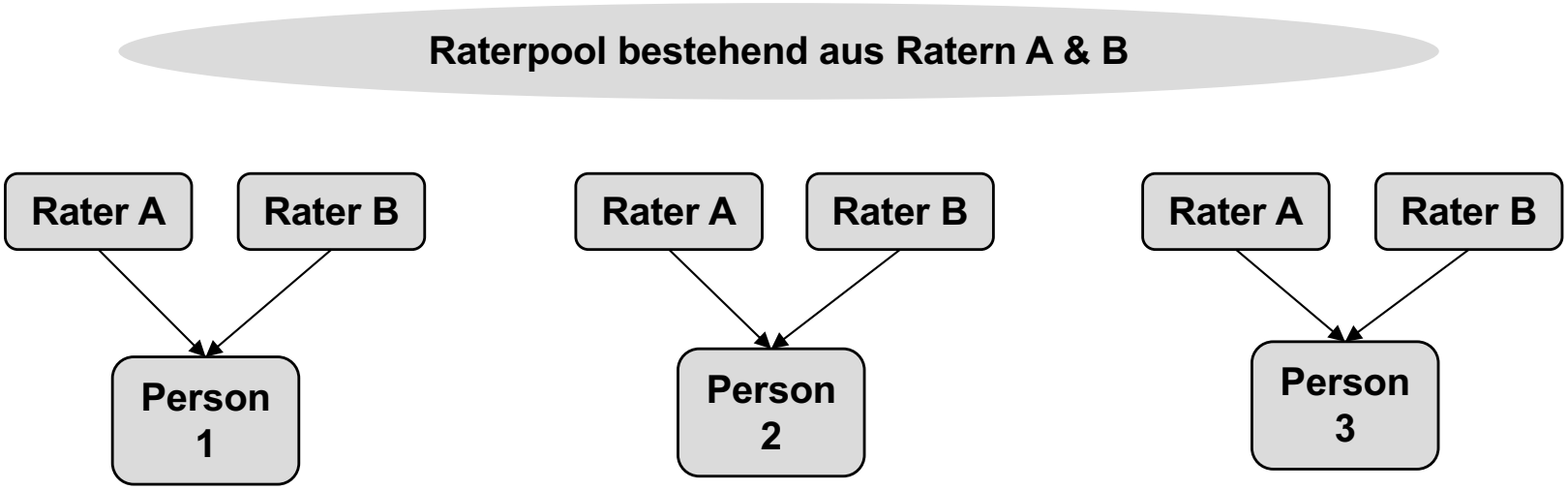
Unterschiede zu Modell 1:

- Modellierung (d.h. Quantifizierung) *und* Berücksichtigung des Fehlers, der durch unterschiedliche Strenge entsteht
- Für die Berechnung muss es eine Überschneidung zwischen Personen und Ratern geben. Im Idealfall wird jede Person von jedem Rater beurteilt (siehe Exkurs 1 im Anhang).
- Genauere (d.h. bessere) Schätzung der Varianz der wahren Unterschiede im Merkmal (vgl. Wirtz & Caspar, S. 181):
  - Wenn die Daten vorliegen, um eine ICC2 zu berechnen, dann ist diese der ICC1 vorzuziehen
  - Wenn systematische Ratervarianz vorhanden ist, dann unterschätzt die ICC1 tendenziell die Reliabilität

Modell „ICC3“: ICC 2-faktoriell, fixed, justiert

Anwendungs-Szenario:

- Es gibt einen Raterpool von *n* Ratern
- Jede Person wird von *allen diesen Ratern* beobachtet, diese Rater sind die einzigen Rater von Interesse (→ fixed), z.B.:



Prototypisches Beispiel:

- In einem Unternehmen werden Bewerberinnen in Assessment-Centern beurteilt
- Im Unternehmen gibt es 2 Leute, die eine Schulung für die Beurteilung in diesem Assessment-Center absolviert haben
- Innerhalb eines ACs wird jede Bewerberin von diesen gleichen 2 Ratern beurteilt
- Zwischen den ACs wechseln die Rater nicht, die berechneten Beurteilungsübereinstimmungsmaße gelten also nur für diese 2 Rater

Eigenschaften:

- Varianz zwischen Ratern (d.h. Unterschiede in der Strenge) wird modelliert → 2-faktoriell
- Eine Generalisierung auf eine Population von Ratern ist *nicht* vorgesehen → fixed
- Interpretation: Wie gut stimmen die Rater *relativ* überein? → justiert (Konsistenz)

Varianten:

- Interessiert die Reliabilität eines Urteils → single
- Interessiert die Reliabilität des mittleren Urteils → average

Unterschied zu **Modell 1**:

- *Modellierung (d.h. Quantifizierung)* und *keine Berücksichtigung* des Fehlers, der durch unterschiedliche Strenge entsteht

Unterschiede zu **Modell 2**:

- *Modellierung (d.h. Quantifizierung)*, aber *keine Berücksichtigung* des Fehlers, der durch unterschiedliche Strenge entsteht
- In Modell 3 wird nur über die Rater eine Aussage gemacht, von denen auch Daten vorliegen

Die ICC-Modelle im Verhältnis zueinander

In der Regel gilt: ICC3 > ICC2 > ICC1

- Die Unterschiede zwischen den ICCs sind umso stärker ausgeprägt, je stärker sich die Rater-Mittelwerte unterscheiden (d.h. desto mehr sich die Rater in ihrer Strenge unterscheiden)
- Der Vergleich der Maße kann wertvolle Hinweise geben
  - für nötige Schulungsmaßnahmen: Muss ich am grundlegenden Verständnis der Beurteilungsskalen arbeiten (um die Konsistenz zu erhöhen), oder „nur“ die Strenge kalibrieren?
  - für mögliche Probleme bei Entscheidungsmodellen: Macht es Sinn einen Cutoff mit absoluten Werten heranzuziehen, oder macht ggf. eine Quote mehr Sinn?

Die ICC-Modelle im Verhältnis zueinander

In der Regel gilt: $ICC_{average} > ICC_{single}$

- Die Mittelung von Ratings reduziert generell Fehler bei der Beurteilung
- Die Unterschiede sind u.a. umso stärker ausgeprägt, je mehr Rater in das average-Modell einfließen
- Praktisch interessant ist hierbei wie groß der Unterschied zwischen den Maßen ist
- Exkurs: Die $ICC_{average}$ basiert auf der Annahme von parallelen Ratern. Nur im (essentiell) parallelen Modell erhöht sich die Reliabilität des Ratermittelwerts zwangsläufig mit der Anzahl der Rater. Gilt eigentlich ein weniger strenges Modell (z.B. τ-kongenerisch), kann die Reliabilität des Ratermittelwerts bei Hinzunahme unreliabler Rater auch abnehmen.

Was mache ich, wenn ich das mittlere Urteil für k Rater wissen möchte, aber Daten von m Ratern vorliegen habe?

Möglichkeit 1: Anwendung der Spearman-Brown-Formel (siehe Testtheorie)

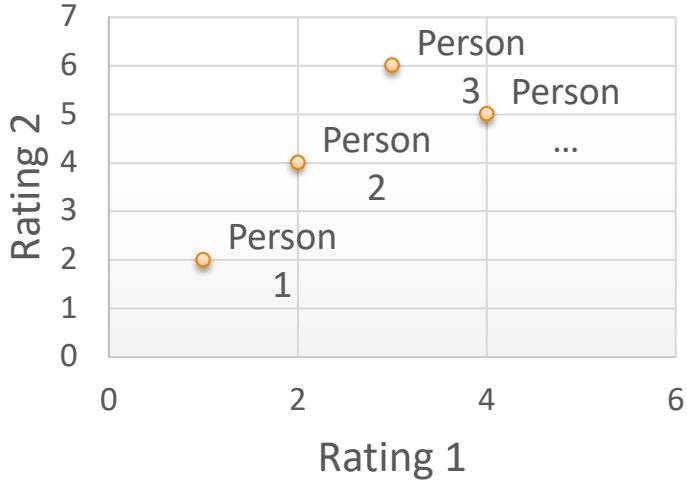
- Aus der Reliabilitätschätzung des Einzelurteils wird die Reliabilität für beliebig viele k Rater geschätzt:
  - $ICC_{average} = \frac{k \cdot ICC_{single}}{1+(k-1) \cdot ICC_{single}}$
- Es ist sogar möglich, basierend auf der Reliabilitätsschätzung des Einzelurteils zu berechnen, wie viele Rater k man benötigen würden, um eine bestimmte gewünschte Reliabilität $ICC_{average}$ des mittleren Urteils zu erreichen:
  - $k = \frac{ICC_{average} \cdot (1-ICC_{single})}{ICC_{single} \cdot (1-ICC_{average})}$

Was mache ich, wenn ich das mittlere Urteil für k Rater wissen möchte, aber Daten von m Ratern vorliegen habe?

<u>Möglichkeit 2: pragmatisch, aber eher nicht zu empfehlen:</u>

- Reduktion der Anzahl der Rater m im Datensatz auf k zufällige Rater und anschließende Anwendung der ICC()-Funktion auf dem reduzierten Datensatz, also z.B. auf k = 3 zufällige Rater (aus m = 5), um die Schätzung für ICC(2,3) zu bekommen
- Wenn man sich für Reliabilität im ICC-Modell 3 interessiert (→ "fixed"), sollte man die Daten entsprechend auf die festen (konkreten) Rater reduzieren, die in Zukunft eingesetzt werden sollen, um die richtige Schätzung für ICC(3,3) zu erhalten

Exkurs: Woher kommt der Name „Intraklassenkorrelation“?



Jede ICC kann als Schätzwert
für die Korrelation zwischen zwei Ratings
der gleichen Personen in der Population
interpretiert werden:

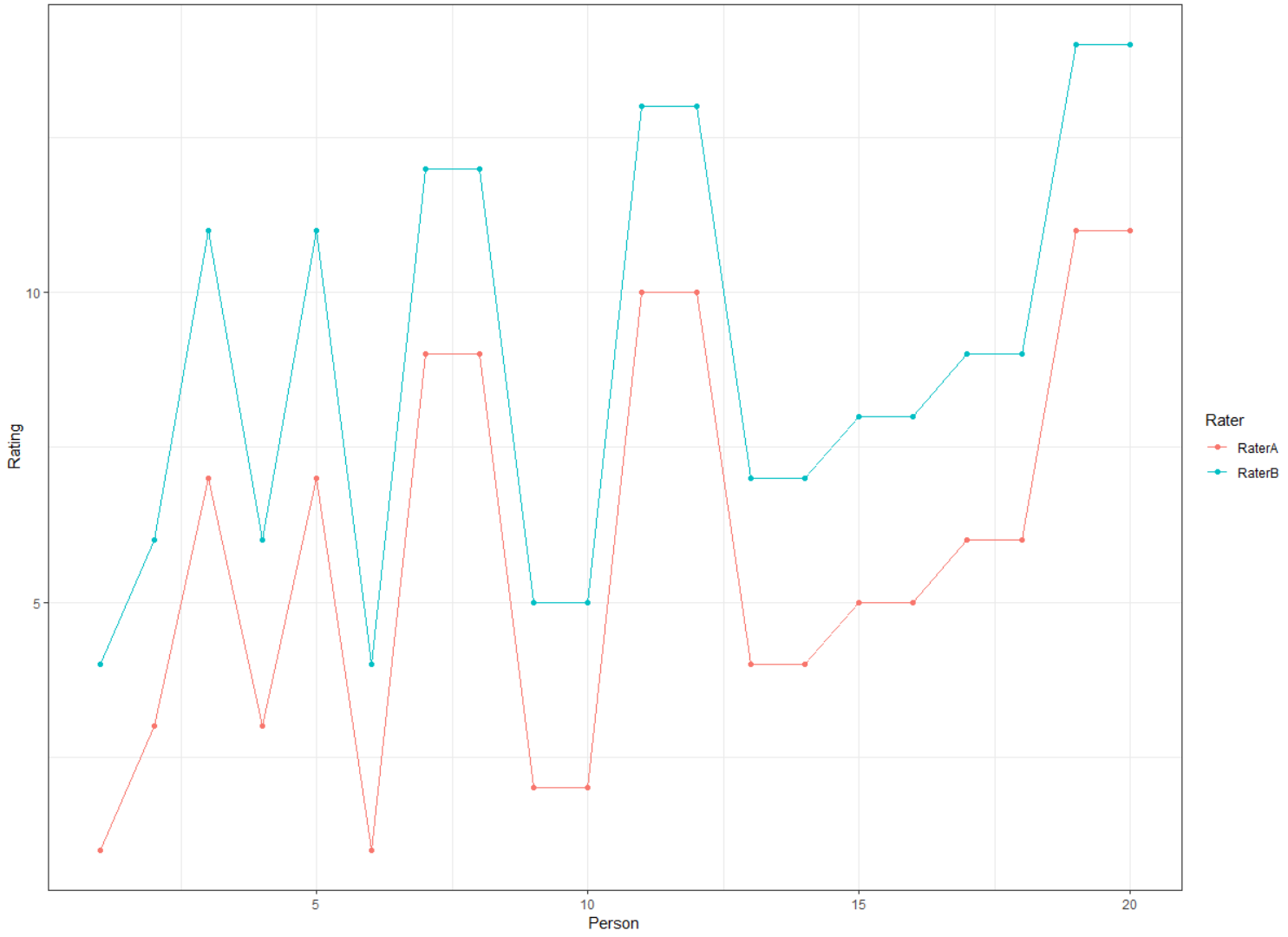
- ICC(1,1): Für jede Person werden die beiden Rater zufällig gezogen. (In Modell 1 können die Rater ohnehin nicht identifiziert werden)
- ICC(2,1): Für jede Person werden die beiden Rater zufällig gezogen. Jede Person wird also von zwei neuen Ratern beurteilt.
- ICC(3,1): Es werden zwei Rater (Rater 1 und Rater 2) zufällig gezogen und diese beiden Rater beurteilen alle Personen.

Für die average ICCs bezieht sich die Korrelation analog auf zwei mittlere Ratings (mittleres Rating 1 und 2) der gleichen Personen.

ICCs in R → Bezug ICC zur Korrelation

data-Objekt

Person	RaterA	RaterB
1	1	4
2	3	6
3	7	11
4	3	6
5	7	11
6	1	4
7	9	12
8	9	12
9	2	5
10	2	5
11	10	13
12	10	13
13	4	7
14	4	7
15	5	8
16	5	8
17	6	9
18	6	9
19	11	14
20	11	14



ICCs in R

data-Objekt
(anders sortiert):

Person	RaterA	RaterB
1	1	4
6	1	4
9	2	5
10	2	5
2	3	6
4	3	6
13	4	7
14	4	7
15	5	8
16	5	8
17	6	9
18	6	9
3	7	11
5	7	11
7	9	12
8	9	12
11	10	13
12	10	13
19	11	14
20	11	14

Code:

```
library(psych)
ICC(data[, c("RaterA", "RaterB")])
```

Output:

```
Call: ICC(x = data)

Intraclass correlation coefficients
                         type  ICC     F df1 df2       p lower bound upper bound
Single_raters_absolute    ICC1 0.65   4.7  19  20 6.2e-04      0.3040        0.84
Single_random_raters      ICC2 0.70 476.0  19  19 1.7e-21     -0.0022        0.93
Single_fixed_raters       ICC3 1.00 476.0  19  19 1.7e-21      0.9894        1.00
Average_raters_absolute  ICC1k 0.78   4.7  19  20 6.2e-04      0.4663        0.91
Average_random_raters    ICC2k 0.82 476.0  19  19 1.7e-21     -0.0043        0.96
Average_fixed_raters     ICC3k 1.00 476.0  19  19 1.7e-21      0.9947        1.00

 Number of subjects = 20     Number of Judges =  2
```

Konfidenzintervalle:

Weitere justierte Maße (bei zwei Ratern)

Für intervallskalierte Daten: Pearson Korrelation

- Die Pearson-Korrelation entspricht dem ICC(3,1) falls Varianzhomogenität zwischen den Ratern vorliegt
- Falls keine Varianzhomogenität vorliegt, ist das ICC-Modell 3 der Pearson-Korrelation vorzuziehen, da die Varianzunterschiede mit berücksichtigt werden

Für ordinalskalierte Daten:

- Spearman Rangkorrelation
- Kendalls Tau

Für alle Maße gilt, wie bei jeder Schätzung:

Je größer die Stichprobengröße, desto präziser ist die Schätzung der Varianzquellen und in der Folge auch die Schätzung der Reliabilität!

Unrepräsentative Stichproben sind natürlich auch hier ein Problem!

7. Zusammenfassung & Beispiele

Zusammenfassung: Anwendung

ICC type	**Description**
ICC(1,1)	Each subject is assessed by a *different set* of *randomly selected* raters, and the reliability is calculated from a single measurement. Uncommonly used in clinical reliability studies.
ICC(1,k)	As above, but reliability is calculated by taking an average of the *k* raters' measurements.
ICC(2,1)	Each subject is measured by each rater, and raters are considered representative of a larger population of similar raters. Reliability calculated from a single measurement.
ICC(2,k)	As above, but reliability is calculated by taking an average of the *k* raters' measurements.
ICC(3,1)	Each subject is assessed by each rater, but the raters are the only raters of interest. Reliability calculated from a single measurement.
ICC(3,k)	As above, but reliability is calculated by taking an average of the *k* raters' measurements.

Zusammenfassung: Anwendung & Interpretation

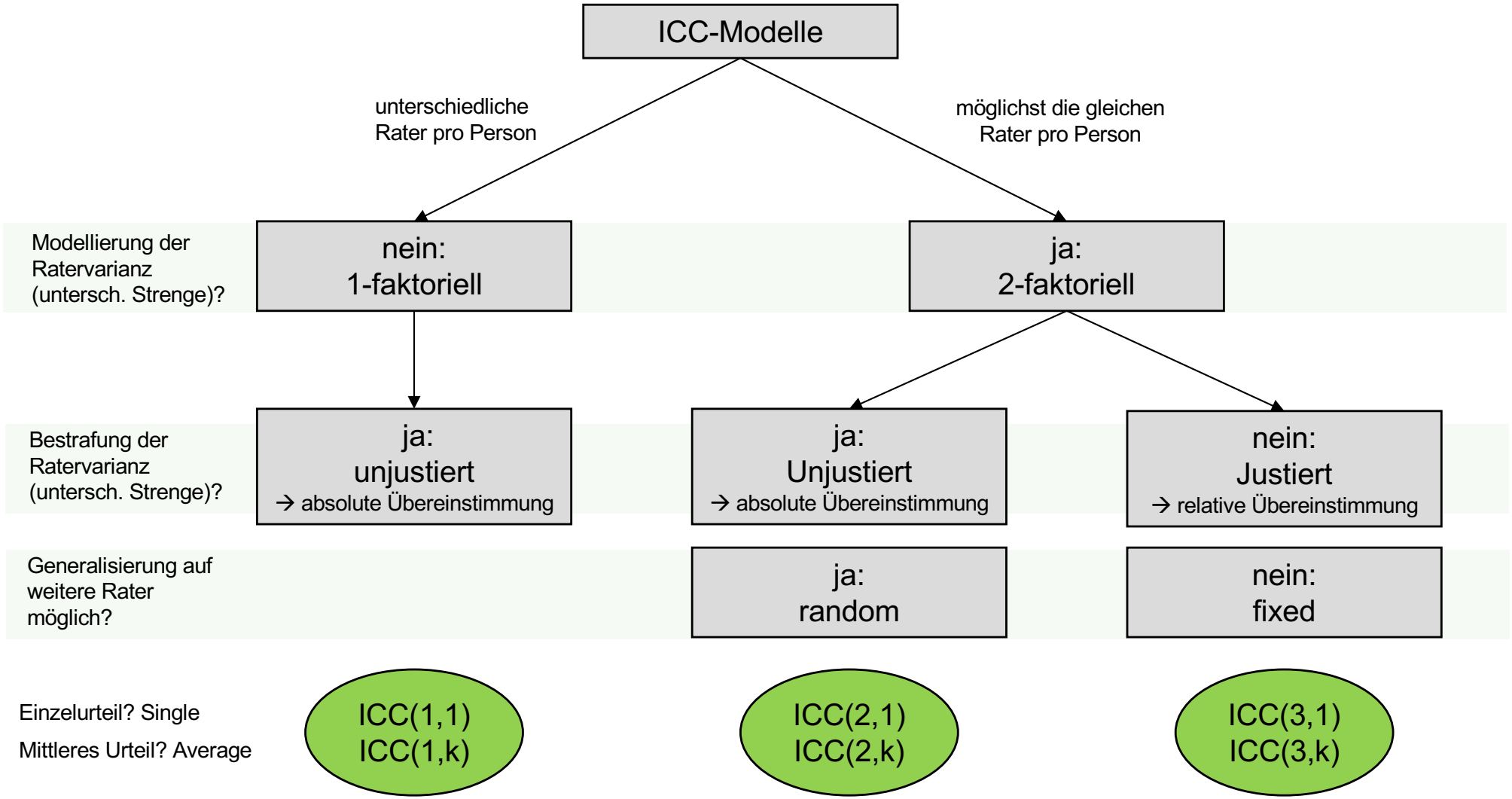
Tabelle 6.4: Überblick über die Entscheidungskriterien bei der Anwendung der drei ICC's.

ICC	Eigenschaften der Raterstichprobe[a)]	Interpretation
$ICC_{unjust.einfakt}$	Die Objekte können jeweils von unterschiedlichen Ratern beurteilt worden sein.	Die absoluten Skalenwerte werden unabhängig vom jeweiligen Rater interpretiert.
ICC_{unjust}	Alle Objekte müssen von denselben Ratern beurteilt worden sein. Nur wenn die Raterstichprobe eine zufällige Auswahl der Rater aus der Population darstellt, ist die ICC_{unjust} ein Reliabilitätsmaß.	
ICC_{just}	Alle Objekte müssen von allen Ratern beurteilt worden sein. Nur wenn die Reliabilitätsaussage ausschließlich für die Rater, die tatsächlich der Untersuchungsstichprobe angehören, gelten soll, ist die ICC_{just} ein Reliabilitätsmaß.	Die Skalenwerte werden relativ zu den übrigen Werten, die der jeweilige Rater vergibt, interpretiert.

a) Die Objekte müssen stets eine Zufallsstichprobe darstellen

Aus Wirtz & Caspar, S.190

Zusammenfassung & Beispiele



Beispiel

In einer Firma gibt es 4 Rater, die im Rahmen von Assessment-Centern immer wieder zu zweit Kandidatinnen beurteilen.

- Alle Rater haben in einem justierten Übereinstimmungsmaß mit allen anderen Ratern einen Wert von r (Pearson-Korrelation) > .75
- Die Beobachterübereinstimmung eines einzelnen Urteils in einem unjustierten Übereinstimmungsmaß beträgt ICC(1,1) = .45
- Die Beobachterübereinstimmung des gemittelten Urteils in einem unjustierten Übereinstimmungsmaß beträgt ICC(1,2) = .60

Was können wir daraus folgern?

- Die Rater sind unterschiedlich streng, bringen die Personen aber (einigermaßen) in die richtige Rangreihe
  - Wenn die absoluten Werte interessieren, sollten die Rater entsprechend geschult werden
  - Wenn nur die Rangfolge interessiert (z.B. bei fixer Annahmequote), könnte man sich ggf. zufriedengeben

Beispiel

In einer Firma gibt es 4 Rater, die im Rahmen von Assessment-Centern immer wieder zu zweit Kandidatinnen beurteilen.

- r (Pearson-Korrelation) > .75
- ICC(1,1) = .45
- ICC(1,2) = .60

Angenommen, uns interessiert die absolute Beurteilung, und es wird vorgeschlagen, zukünftige Kandidatinnen nur noch durch ein Einzelurteil auszuwählen. Ist das eine gute Idee?

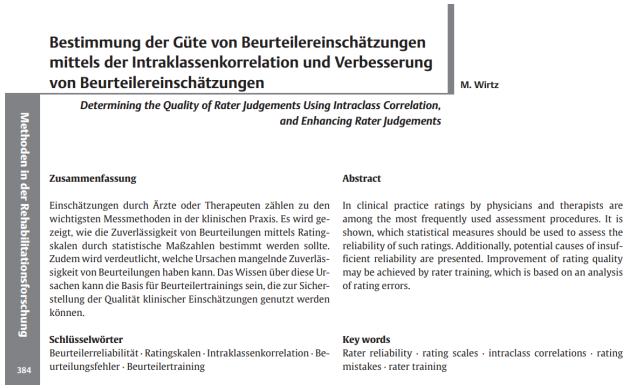
- D.h. uns interessieren nur noch die Werte der beiden unjustierten ICCs
- Die Reliabilität eines Einzelurteils ist mit .45 sehr viel kleiner als die Reliabilität des gemittelten Urteils mit .60 → keine gute Idee

Das war‘s schon fast für heute …

- *Ausblick:* In der nächsten Vorlesung lernen wir mit den Interviews das dritte Instrument zur Erhebung diagnostisch relevanter Informationen in der Psychologie kennen.
- Aber zuerst: Gibt es offene Fragen zur heutigen Vorlesung?
- Achtung: Die Sitzung nächste Woche entfällt wegen des Feiertags!

- Weiterführende Literatur:

Wirtz, M. & Caspar, F. (2004). Beobachterübereinstimmung (ab S. 47) Kapitel 4.1.3, 4.1.4 und Kapitel 6. Weinheim: Juventa.



Bestimmung der Güte von Beurteilereinschätzungen mittels der Intraklassenkorrelation und Verbesserung von Beurteilereinschätzungen

M. Wirtz

Determining the Quality of Rater Judgements Using Intraclass Correlation, and Enhancing Rater Judgements

Methoden in der Rehabilitationsforschung

384

Zusammenfassung

Einschätzungen durch Ärzte oder Therapeuten zählen zu den wichtigsten Messmethoden in der klinischen Praxis. Es wird gezeigt, wie die Zuverlässigkeit von Beurteilungen mittels Ratingskalen durch statistische Maßzahlen bestimmt werden sollte. Zudem wird verdeutlicht, welche Ursachen mangelnde Zuverlässigkeit von Beurteilungen haben kann. Das Wissen über diese Ursachen kann die Basis für Beurteilertrainings sein, die zur Sicherstellung der Qualität klinischer Einschätzungen genutzt werden können.

Schlüsselwörter

Beurteilerreliabilität · Ratingskalen · Intraklassenkorrelation · Beurteilungsfehler · Beurteilertraining

Abstract

In clinical practice ratings by physicians and therapists are among the most frequently used assessment procedures. It is shown, which statistical measures should be used to assess the reliability of such ratings. Additionally, potential causes of insufficient reliability are presented. Improvement of rating quality may be achieved by rater training, which is based on an analysis of rating errors.

Key words

Rater reliability · rating scales · intraclass correlations · rating mistakes · rater training

Wirtz, M. (2004). Bestimmung der Güte von Beurteilereinschätzungen mittels der Intraklassenkorrelation und Verbesserung von Beurteilereinschätzungen. *Die Rehabilitation, 4*, 384-389.

Exkurs 1: Muss in 2-faktoriellen Modellen jeder Rater jede Person gesehen haben?

Hintergrund: ANOVA Berechnung

Die Berechnung der ICCs erfolgte früher mit Hilfe der Varianzanalyse (ANOVA), und diese benötigt generell **„balancierte“ Daten:**

- D.h. alle Gruppen die als Faktor in das Modell eingehen, müssen die gleiche Anzahl an Beobachtungen haben
- D.h. bei einem 2-faktoriellen Design bräuchte es jeweils pro Person *und* pro Rater die gleiche Anzahl an Ratings
- Zum Vergleich: Bei einem 1-faktoriellen Design braucht es nur die gleiche Anzahl an Ratings pro Person

Vorteile LMMs gegenüber ANOVA

- Heute werden die ICCs allerdings meist mithilfe von **gemischten linearen Modellen** (LMMs) geschätzt, die keine balancierten Daten voraussetzen
- Dieser Ansatz wird auch von der „ICC()“-Funktion aus dem psych Paket in R standardmäßig benutzt
  - In der Konsequenz erhalten Sie in R auch bei fehlenden Werten für manche Personen von manchen Ratern Schätzungen für das ICC Modell 2 und ICC Modell 3 (solange eine Überlappung von Personen und Rater vorliegt)
  - Bei fehlenden Werten ist jedoch nicht mehr garantiert, dass die ICCs sinnvoll interpretiert werden können, da die Varianz zwischen den Ratermittelwerten nicht mehr eindeutig als Effekt unterschiedlicher Strenge interpretierbar ist (sondern auch auf unterschiedliche mittlere wahre Merkmalsausprägungen der verschiedenen Personen zurückzuführen sein könnte, die die Rater jeweils bewertet haben).

Datenbeispiele

Auch für diese extremeren Beispieldaten erhalten wir in R Schätzungen für alle ICC-Modelle, obwohl nicht jeder Rater jede Person beobachtet hat:

	RaterA	RaterB	RaterC	RaterD
1	1	4	NA	NA
2	3	6	NA	NA
3	7	11	NA	NA
4	3	6	NA	NA
5	7	11	NA	NA
6	1	4	NA	NA
7	NA	NA	8	12
8	NA	NA	9	12
9	NA	NA	9	5
10	NA	NA	10	5
11	NA	NA	10	13
12	NA	NA	11	13
13	NA	NA	12	14

RaterA	RaterB	RaterC	RaterD	RaterE	RaterF
1	4	NA	NA	NA	NA
2	NA	4	NA	NA	NA
4	NA	NA	2	NA	NA
NA	NA	2	3	NA	NA
NA	NA	NA	NA	4	5
NA	3	NA	NA	NA	1
NA	2	NA	NA	1	NA

Können wir davon ausgehen, dass unterschiedliche Ratermittelwerte in diesen Daten unterschiedliche Strenge widerspiegeln?
Nur unter bestimmten Annahmen!

Mögliche Probleme

Würde der klassische varianzanalytische Ansatz verwendet werden, müssten bei einem zweifaktoriellen Modell Personen mit fehlenden Werten von Ratern komplett aus der Analyse ausgeschlossen werden

- Das heißt, dass auch andere ICC-Werte resultieren würden (oder sie nicht berechnet werden könnten)!
- Zum Testen in R: Argument „lmer = FALSE“ in der ICC-Funktion
  → wirft Fehler wenn missings vorhanden sind

Daher: Vorsicht bei der Interpretation, wenn Sie in R ein ICC-Modell 2 oder 3 auf Daten mit fehlenden Werten berechnen

In der Praxis

Je größer die Überlappung der Personen ist, die von den verschiedenen Ratern bewertet wurden, desto unkritischer ist es die ICC2 und ICC3 auch in diesen Fällen zu interpretieren:

Keine Überlappung:

	RaterA	RaterB	RaterC	RaterD
1	1	4	NA	NA
2	3	6	NA	NA
3	7	11	NA	NA
4	3	6	NA	NA
5	7	11	NA	NA
6	1	4	NA	NA
7	NA	NA	8	12
8	NA	NA	9	12
9	NA	NA	9	5
10	NA	NA	10	5
11	NA	NA	10	13
12	NA	NA	11	13
13	NA	NA	12	14

Große Überlappung:

	RaterA	RaterB	RaterC	RaterD
1	1	4	NA	NA
2	3	6	3	1
3	7	11	5	1
4	3	6	6	4
5	7	11	8	5
6	1	4	9	6
7	3	5	8	12
8	4	5	9	12
9	6	3	9	5
10	1	2	10	5
11	1	1	10	13
12	2	7	11	13
13	NA	NA	12	14

Exkurs 2: Einordnung der ICC Modelle in der Testtheorie

ICC Modell 1

Einfaktorielles Modell:

In der Literatur zu gemischten linearen Modellen (Vorlesung im Master) wird das einfaktorielle Modell häufig folgendermaßen notiert:

$$\underbrace{X_{ip}}_{X_i} = \underbrace{\gamma + u_p}_{\theta_{Person}} + \underbrace{\varepsilon_{ip}}_{\varepsilon_i} \text{ mit } u_p \sim N(0, \underbrace{\sigma^2_{Person}}_{VAR(\theta)}),\ \varepsilon_{ip} \sim N(0, \underbrace{\sigma^{2*}_{\varepsilon}}_{VAR(\varepsilon_i)^*})$$

Wenn wir die Schreibweise aus der Testtheorie-Vorlesung anwenden, erkennen wir, dass es sich hier um ein **paralleles Testmodell** handelt, bei dem die Items i durch Rater ersetzt wurden:

$X_i = \tau_i + \varepsilon_i = \theta_{Person} + \varepsilon_i$ mit konstanter Fehlervarianz $VAR(\varepsilon_i)^*$

Erkenntnis: Es gelten die aus der Testtheorie bekannten Folgerungen und Reliabilitätsschätzungen für das parallele Modell.

ICC Modell 1

Folgerung für die Varianzen im einfaktoriellen Modell:

$$VAR(X_i) = VAR(\tau_i) + VAR(\varepsilon_i)^* = VAR(\theta) + VAR(\varepsilon_i)^*$$

Intraklassenkorrelation einfaktoriell, unjustiert, single:

$$ICC_{1,un,single} = \frac{VAR(\theta)}{VAR(\theta) + VAR(\varepsilon_i)^*}$$

→ Geschätzter Anteil an der Gesamtvarianz, der **nicht** auf unterschiedliche Beurteilungen zurückzuführen ist

→ Wie reliabel ist ein **Einzelurteil**?

→ Da $VAR(rater_i)$ nicht geschätzt wird, ist unterschiedliche Strenge und mangelnde Konsistenz in $VAR(\varepsilon_i)^*$ vermischt

Die $ICC_{1,un,single}$ wird in der Klassifikation von Shrout & Fleiss als $ICC(1,1)$ bezeichnet. Sie entspricht der aus der Testtheorie bekannten Itemreliabilität für das parallele Modell.

ICC Modell 1

Intraklassenkorrelation einfaktoriell, unjustiert, average:

$$ICC(1,k) = \frac{k \cdot ICC(1,1)}{1 + (k-1) \cdot ICC(1,1)} = \cdots = \frac{VAR(\theta)}{VAR(\theta) + \frac{VAR(\varepsilon_i)^*}{k}}$$

→ Geschätzter Anteil an der Varianz der Mittelwerte, der **nicht** auf unterschiedliche Beurteilungen zurückzuführen ist

→ Wie reliabel ist das **mittlere Urteil**?

Die $ICC_{1,un,average}$ wird in der Klassifikation von Shrout & Fleiss als $ICC(1,k)$ bezeichnet. Sie entspricht der aus der Testtheorie bekannten Reliabilität für den Itemmittelwert im parallelen Modell und berechnet sich mithilfe der Spearman-Brown Formel. Der Itemmittelwert entspricht hier dem Ratermittelwert aus k Ratern.

ICC Modell 2

Zweifaktorielles Modell:

In der Literatur zu gemischten linearen Modellen (Vorlesung im Master) wird das zweifaktorielle Modell häufig folgendermaßen notiert:

$$\underbrace{X_{ip}}_{X_i} = \underbrace{\gamma + u_p}_{\theta_{Person}} + \underbrace{u_i}_{\sigma_i} + \underbrace{\varepsilon_{ip}}_{\varepsilon_i} \text{ mit } u_p \sim N(0, \underbrace{\sigma^2_{Person}}_{VAR(\theta)}),\ u_i \sim N(0, \underbrace{\sigma^2_{Rater}}_{VAR(rater_i)}),\ \varepsilon_{ip} \sim N(0, \underbrace{\sigma^2_{\varepsilon}}_{VAR(\varepsilon_i)})$$

Wenn wir die Schreibweise aus der Testtheorie anwenden, erkennen wir, dass es sich hier um ein **essentiell paralleles Testmodell** handelt, bei dem die Items i durch Rater ersetzt werden:

Exkurs: Die Rater sind hier anders als in Testtheorie „zufällig“, d.h. die Parameter σ_i werden nicht direkt geschätzt, sondern nur deren Varianz $VAR(\sigma_i) \neq 0$

$X_i = \tau_i + \varepsilon_i = \sigma_i + \theta_{Person} + \varepsilon_i$ mit konstanter Fehlervarianz $VAR(\varepsilon_i)$

Erkenntnis: Es gelten die aus der Testtheorie bekannten Folgerungen und Reliabilitätsschätzungen für das essentiell parallele Modell.

ICC Modell 2

Folgerung für die Varianzen im zweifaktoriellen Modell:

$$VAR(X_i) = \tau_i + \varepsilon_i = VAR(\theta) + VAR(rater_i) + VAR(\varepsilon_i)$$

Intraklassenkorrelation zweifaktoriell, unjustiert, single:

$$ICC_{2,un,single} = \frac{VAR(\theta)}{VAR(\theta) + VAR(rater_i) + VAR(\varepsilon_i)}$$

Die $ICC_{2,un,single}$ wird in der Klassifikation von Shrout & Fleiss als $ICC(2,1)$ bezeichnet. Sie entspricht einer bestimmten Itemreliabilität für das essentiell parallele Modell. Aber weil hier die Rater als „zufällig“ angenommen werden (und damit $VAR(rater_i) \neq 0$), ist diese Reliabilität *nicht* identisch mit der aus der Testtheorie bekannten Itemreliabilität (diese kommt gleich noch).

ICC Modell 2

Intraklassenkorrelation zweifaktoriell, unjustiert, average:

$$ICC(2,k) = \frac{k \cdot ICC(2,1)}{1 + (k-1) \cdot ICC(2,1)} = \cdots = \frac{VAR(\theta)}{VAR(\theta) + \frac{VAR(rater_i) + VAR(\varepsilon_i)}{k}}$$

Die $ICC_{2,un,average}$ wird in der Klassifikation von Shrout & Fleiss als $ICC(2,k)$ bezeichnet. Sie entspricht einer Reliabilität des Itemmittelwerts im essentiell parallelen Modell und berechnet sich mithilfe der Spearman-Brown Formel. Der Itemmittelwert entspricht hier dem Ratermittelwert aus k Ratern. Weil die $ICC(2,1)$ nicht genau der aus der Testtheorie bekannten Itemreliabilität entspricht, ist auch diese Reliabilität des Mittelwerts *nicht* identisch mit der aus der Testtheorie bekannten Reliabilität des Itemmittelwerts (diese kommt gleich noch).

ICC Modell 3

Folgerung für die Varianzen im zweifaktoriellen Modell:

$$VAR(X_i) = \tau_i + \varepsilon_i = VAR(\theta) + VAR(rater_i) + VAR(\varepsilon_i)$$

Intraklassenkorrelation zweifaktoriell, justiert, single:

→ Die Rater werden als Faktor im Modell geschätzt, aber nicht als Fehler berücksichtigt, $VAR(rater_i)$ wird für die Berechnung von der Gesamtvarianz abgezogen:

$$ICC_{2,jus,single} = \frac{VAR(\theta)}{VAR(\theta) + VAR(\varepsilon_i) + VAR(rater_i) - VAR(rater_i)}$$

Gesamtvarianz

Restliche Fehlervarianz, nachdem für Rater kontrolliert wurde

= 0

ICC Modell 3

$$ICC_{2,jus,single} = \frac{VAR(\theta)}{VAR(\theta) + VAR(\varepsilon_i)}$$

Die $ICC_{2,jus,single}$ wird in der Klassifikation von Shrout & Fleiss als $ICC(3,1)$ bezeichnet. Sie entspricht der aus der Testtheorie bekannten Itemreliabilität für das essentiell parallele Modell.

→ Modellierung der Rater, *aber keine* Berücksichtigung des Fehlers, der durch unterschiedliche Strenge entsteht

→ Es wird nur über die Rater eine Aussage gemacht, von denen Daten vorliegen

Hinweis: Je nach Notation scheinbar gleiche Formel wie bei $ICC(1,1)$, jedoch sind die Varianzen hier aus dem zweifaktoriellen Modell (siehe Definition und Modellgleichung).

ICC Modell 3

Intraklassenkorrelation zweifaktoriell, justiert, average:

$$ICC(3,k) = \frac{k \cdot ICC(3,1)}{1 + (k-1) \cdot ICC(3,1)} = \cdots = \frac{VAR(\theta)}{VAR(\theta) + \frac{VAR(\varepsilon_i)}{k}}$$

Die $ICC_{3,jus,average}$ wird in der Klassifikation von Shrout & Fleiss als $ICC(3,k)$ bezeichnet. Sie entspricht der aus der Testtheorie bekannten Reliabilität für den Itemmittelwert im essentiell parallelen Modell und berechnet sich mithilfe der Spearman-Brown Formel. Der Itemmittelwert entspricht hier dem Ratermittelwert aus k Ratern.

Hinweis 1: Man kann zeigen, dass die $ICC(3,k)$ Cronbach‘s α entspricht.

Hinweis 1: Je nach Notation scheinbar gleiche Formel wie bei $ICC(1,k)$, jedoch sind die Varianzen hier aus dem zweifaktoriellen Modell (siehe Definition und Modellgleichung).