Grundlagen der Diagnostik

Lerneinheit 6

Beobachter- und Beurteilerübereinstimmung III

Wahr-oder-Falsch Aussagen zur letzten Sitzung

1. Pearson Korrelationen eignen sich *per se* nicht als Beurteilungsmaß für intervallskalierte Daten.
2. Bei Intra-Klassen-Korrelationen werden bis zu drei Quellen von Varianzen berücksichtigt.
3. Wenn man für bestimmte Rater (ohne Generalisierung) die ICC wissen möchte, braucht man eine Single ICC (steht für single raters).
4. Average ICC sind i.d.R. höher als Single ICC.
5. Die ICC (3,1; d.h. 2-faktoriell, justiert, fixed, single) ist *immer* die konservativste und damit beste Reliabilitätsschätzung im Kontext intervallskalierter Daten.

1. ICC Modelle (Shrout & Fleiss, 1979)

2. Zusammenfassung & Beispiele

Lernziele

1. ICC Modelle (Shrout & Fleiss, 1979)

Rückblick: 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 Modelle (Shrout & Fleiss, 1979)	**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 der der Mittelwert gebildet wurde
z.B. bei 3 Ratern und Modell 1: ICC(1,3)

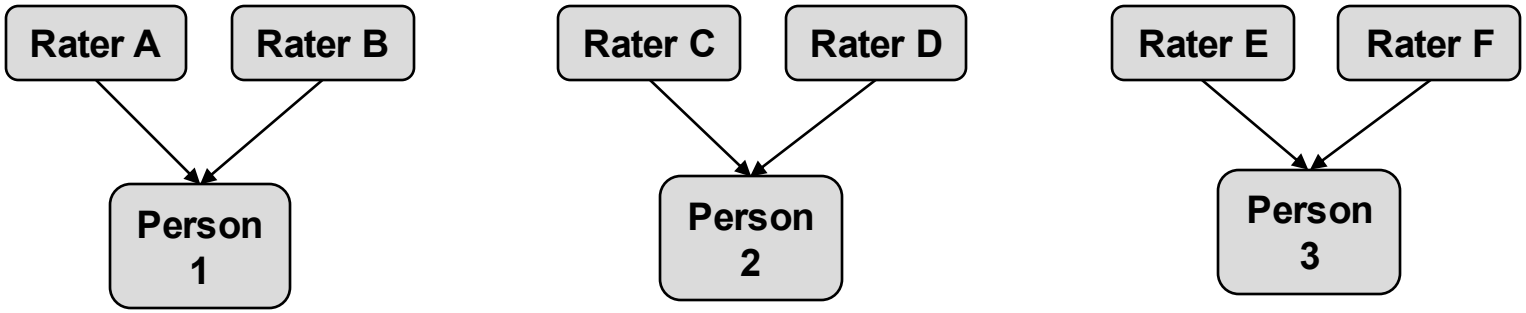
Rückblick: Schätzung der verschiedenen Varianzquellen

- Die Varianz der wahren Unterschiede im Merkmal $VAR(\theta)$, Ratervarianz $VAR(rater_i)$ und Fehlervarianz $VAR(\varepsilon_i)$ sind nicht direkt 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** (LMMs) verwendet.
- Im LMM entsprechen die Varianzen jeweils einem Modellparameter. Damit erhalten wir direkt die folgenden Schätzwerte:
  $\widehat{VAR(\theta)}$, $\widehat{VAR(rater_i)}$ und $\widehat{VAR(\varepsilon_i)}$

Modell „ICC1“: ICC 1-faktoriell & unjustiert

Anwendungs-Szenario:

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

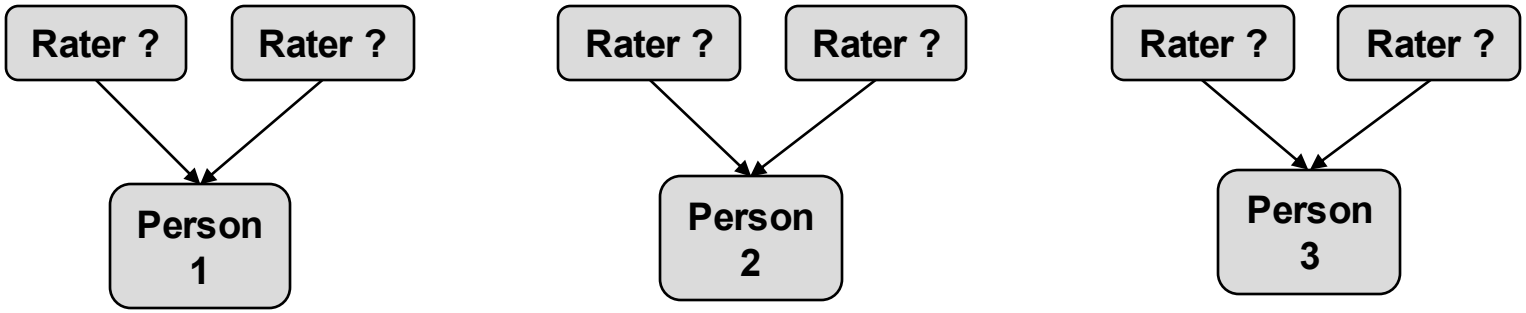


Hinweis: Das ICC1-Modell kann auf verschiedene Datensituationen angewendet werden, wird aber typischerweise für die obige Situation angewandt, da bei Vorliegen von Ratings aller Rater für alle Personen die anderen ICC-Modelle besser geeignet sind.

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

Varianten:

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

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.

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.

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

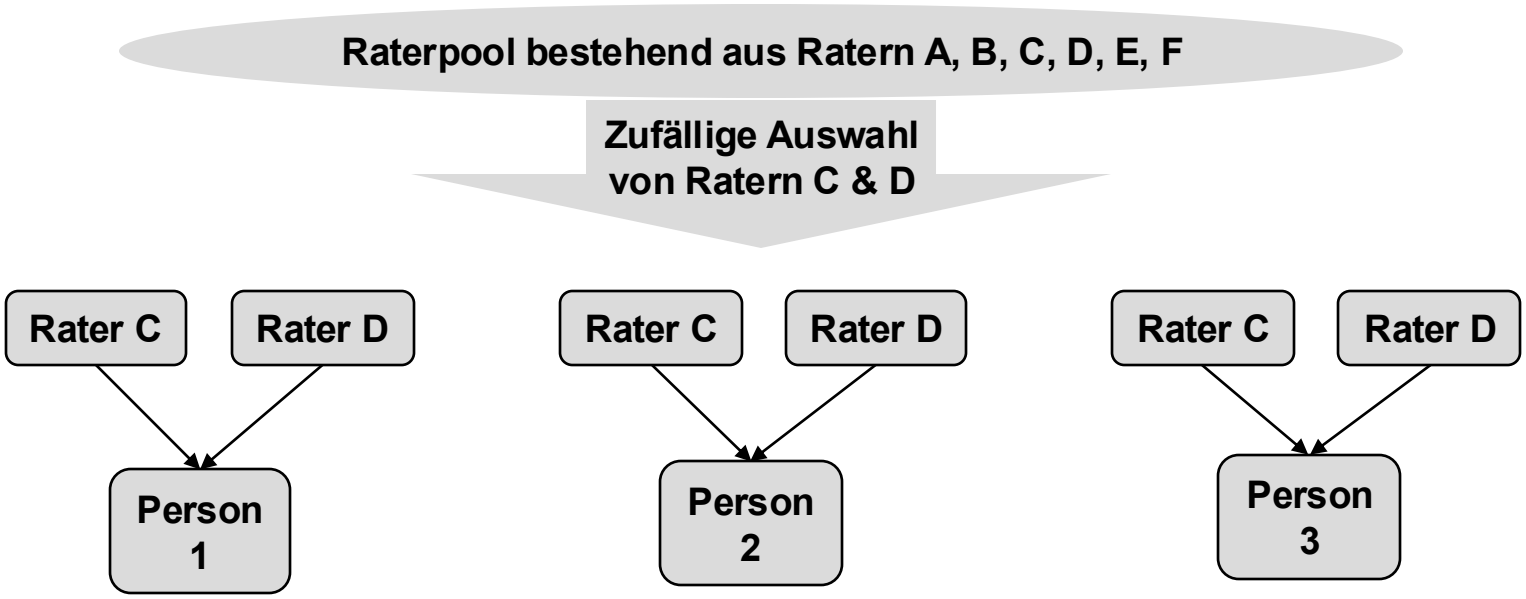
→ Wie reliabel ist das **mittlere Urteil**?

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.

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 wurden (→ 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 & gleiche Strenge)

Varianten:

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

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.

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).

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).

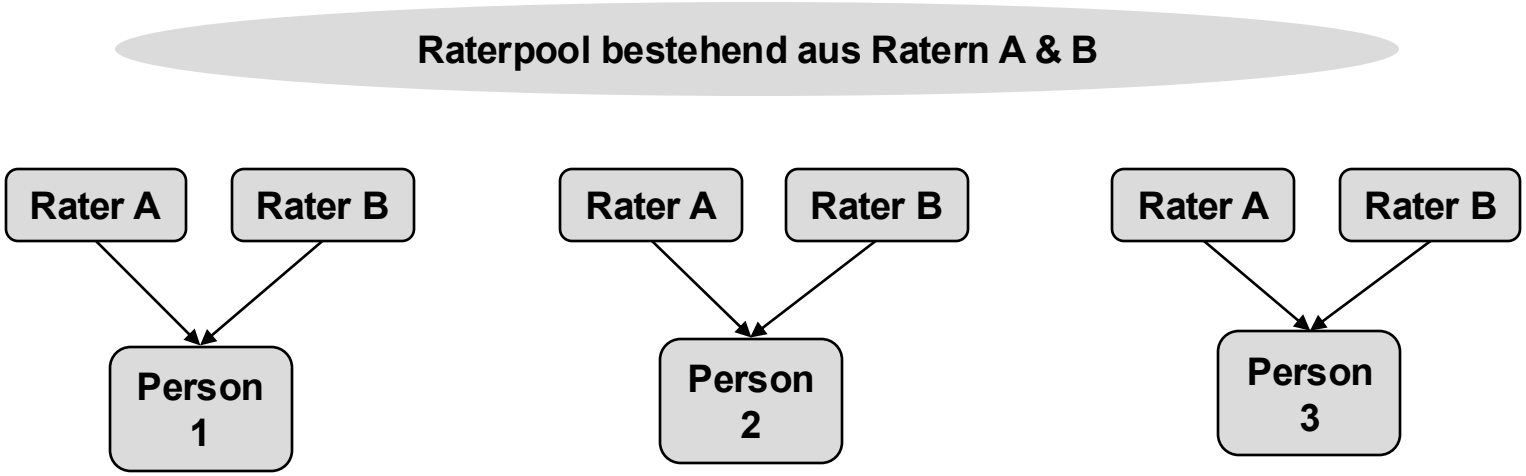
Unterschiede zu Modell 1:

- 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.
- 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

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

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

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).

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).

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
- Achtung: 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 Spearman-Brown-Formel

- Aus der Reliabilitätschätzung des Einzelurteils wird 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

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!

2. 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

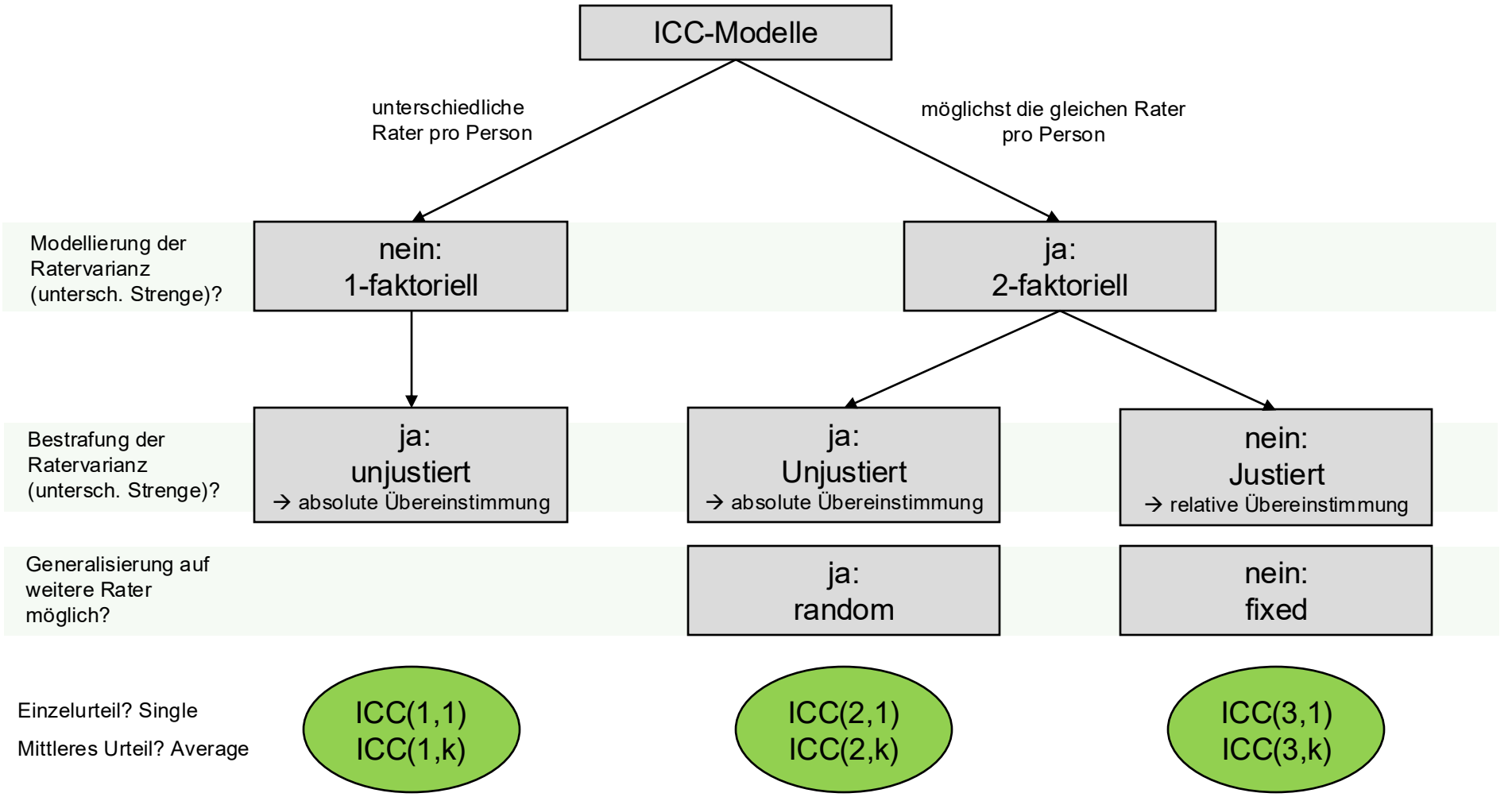
Tabelle 6.4: Überblick über die Entscheidungskriterien bei der Anwendung der drei ICC's.

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 für Assessment Center immer wieder zu zweit Kandidatinnen beurteilen.

- Alle Rater haben in einem justierten Übereinstimmungsmaß mit allen anderen Ratern einen Wert von r (Pearson-Korrelation) > .75
- Aus den Firmen-Daten wurden die Beobachterübereinstimmung eines einzelnen Urteils in einem unjustierten Maß berechnet. Sie beträgt ICC(1,1) = .45
- Die Beobachterübereinstimmung des gemittelten Urteils in einem unjustierten Maß 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 also entsprechend geschult werden

Beispiel

In einer Firma gibt es 4 Rater, die für Assessment Center 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, dass zukünftige Kandidatinnen nur noch durch ein Einzelurteil ausgewählt werden. 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

Fazit

Die ICC-Modelle treffen **unterschiedliche Aussagen** in Bezug auf die Beurteilerübereinstimmung, abhängig davon…

- …ob es ein justiertes oder ein unjustiertes Maß ist (und damit verbunden, ob es ein fixed oder random Maß ist)
- …ob es ein single oder average Maß ist

Die **Datensituation ist relevant** dafür, welches ICC-Modell sinnvoll herangezogen werden kann

- Um die Ratervarianz zu modellieren (→ 2-faktoriell), sollte möglichst jeder Rater jede Person beurteilt haben
- Interessiert man sich für ein unjustiertes Maß, ist die ICC 2 gegenüber der ICC 1 vorzuziehen, wenn die Datensituation das zulässt

Einen Schritt zurück

- Beobachter- und Beurteilerübereinstimmungmaße helfen uns also dabei, die Objektivität bzw. Reliabilität von Verfahren wie z.B. der Verhaltensbeobachtung oder Interviews zu beurteilen
- Sie geben Hinweise darauf, an welchen Stellschrauben man drehen sollte, um die Objektivität bzw. Reliabilität zu verbessern

- Die Hauptgütekriterien verdienen ihren Status als wichtigste Gütekriterien: In der Diagnostik treffen wir Entscheidungen basierend auf den Ergebnissen unserer diagnostischen Verfahren!
- Wir sollten uns auf diese Ergebnisse so gut wie möglich verlassen können, um möglichst wenige diagnostische Fehlentscheidungen zu treffen!