# Einführung in die Test- und Fragebogenkonstruktion

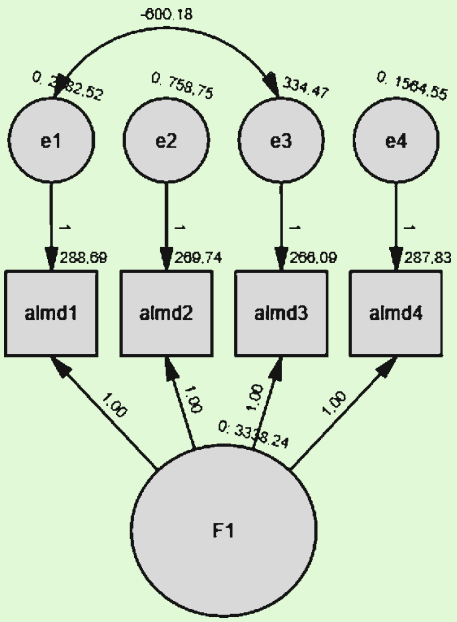
4., korrigierte und erweiterte Auflage

Markus Bühner

Pearson

EXTRAS ONLINE

**Es ergibt sich folgende SPSS-Ausgabe:**




*Reliabilitätsschätzung.* Wir können nun die Reliabilität für den Summenwert schätzen, dafür verwenden wir die folgende Formel (Herleitung siehe ***Anhang I***):

$$\mathrm{REL}\left(\frac{1}{k}\cdot\sum_{i=1}^{k}X_i\right)=\frac{k^2\cdot\mathrm{VAR}\left(\theta\right)}{\left[k^2\cdot\mathrm{VAR}\left(\theta\right)+\sum_{i=1}^{k}{\varepsilon_i}^2+2\cdot\sum_{i\neq j}\mathrm{COV}\left(\varepsilon_i,\varepsilon_j\right)\right]}$$

Wir können die Schätzfunktion für tau-kongenerische Messungen mit korrelierten Fehlern von Raykov nicht einfach übernehmen, da bei einer unstandardisierten latenten Variablen (wie im essenziell tau-äquivalenten Modell) die Varianz der latenten Variablen nicht wie bei Raykovs Omega aus der Schätzfunktion fällt. Wenn wir die Werte von oben einsetzen, erhalten wir die folgende Reliabilitätsschätzung:

$$\mathrm{rel}\left(\frac{1}{k}\cdot\sum_{i=1}^{k}x_i\right)$$

$$=\frac{4^2\cdot 3338.236}{4^2\cdot 3338.236+2082.524+758.752+334.471+1564.553+2\cdot -600.177}$$

$$\mathrm{rel}\left(\frac{1}{k}\cdot\sum_{i=1}^{k}x_i\right)=\frac{53411.78}{53411.78+4740.3-1200.354}=\frac{53411.78}{56951.73}=0.9378$$

*Zusammenfassung.* Endlich haben wir ein Testmodell gefunden, von dem wir ausgehen, dass es die Itemantworten in der Population mit seinen Parametern angemessen beschreibt. Die Verwendung des Summenwerts kann in unserem Beispiel nicht empfohlen werden. Zunächst ist das Modell essenziell tau-äquivalent

und durch die unterschiedliche Itemreliabilität kann nicht jedes Item zur Feststellung der Fähigkeit gleich viel wert sein. Zum anderen haben wir wenige Items und dazu noch zwei korrelierte Fehlervariablen. Das heißt, die Hälfte der Items misst eine zweite latente Variable zusätzlich zur Reaktionsschnelligkeit bzw. Alertness.

*Personenparameter.* Häufig wird argumentiert, dass bei einer hohen Korrelation zwischen Summenwert und Faktorwert (siehe ***Definition 4.1*** und ***Kapitel 6***) die Verwendung von Faktorwerten und von Summenwerten keinen Unterschied ergeben würde. Im Durchschnitt trifft das zu, jedoch nicht bei der Betrachtung von Einzelfällen. Hier spielt auch eine Rolle, ob die Faktorwerte aus einem essenziell tau-äquivalenten Modell mit oder ohne Fehlerkorrelationen geschätzt wurden. Betrachten wir zunächst die Korrelation der Faktorwerte (geschätzt nach der Bartlett-Methode für das essenziell tau-äquivalente Modell ohne korrelierte Fehlervariablen und mit korrelierten Fehlervariablen für unser Datenbeispiel. In R werden über folgenden Befehl die Faktorwerte im Programmpaket `lavaan` ausgegeben:

```
f_almd <- (lavPredict(fit, type="lv", method = "bartlett"))
list(f_almd)
```

Es ergibt sich folgende Ausgabe der unstandardisierten Werte der latenten Variablen in R:

```
            f1
  [1,] -59.259
  [2,] -18.657
  [3,] -60.641❷
  [4,]  46.030
  [5,] -15.178
  [6,]  59.212❶
  [7,]  45.914
  [8,] -11.098
  [9,] -37.416
 [10,] -50.725
...
```

Die Werte stellen Abweichungswerte von der latenten Variablen dar. Hohe positive Werte bedeuten eine langsame Reaktionszeit ❶ (höhere Reaktionszeit als der Mittelwert) und hohe negative Werte eine schnelle Reaktionszeit ❷ (geringere Reaktionszeit als der Mittelwert) auf der latenten Variablen.

Um die Werte mit den Rohwerten vergleichen zu können, müssen wir den Mittelwert der TAP-Tests (almd1 bis almd4) bilden (R: `mean_almd <- ((TAP$almd1 + TAP$almd2 + TAP$almd3 + TAP$almd4)/4)`). Von diesem Mittelwert wird der Mittelwert der Mittelwerte abgezogen (R: `dev_almd_mean <- (mean_almd - (mean(mean_almd)))`), um einen Abweichungswert (`dev_almd_mean`) zu erhalten.

Die Korrelation zwischen den Faktorwerten des essenziell tau-äquivalenten Modells mit dem Mittelwert der TAP-Items beträgt 0.98 (R: `cor(f_almd,`

dev_almd_mean)). Hier wird zunächst deutlich, dass die Werte nicht identisch sind. Betrachten wir nun das Streudiagramm (R: plot(f_almd, dev_almd_mean)) der Faktorwerte des essenziell tau-äquivalenten Modells mit korrelierten Fehlervariablen mit dem Mittelwert der TAP-Items (siehe Abbildung 4.30). In dem Streudiagramm sind auf der *x*-Achse die unstandardisierten Faktorwerte (f_almd) für ein essenziell tau-äquivalentes Modell mit korrelierten Fehlern dargestellt. Auf der *y*-Achse sind die Abweichungswerte (dev_almd_mean) vom Mittelwert (in Millisekunden) des TAP-Gesamttests dargestellt. Jede Person besitzt einen Faktorwert und einen Abweichungswert vom Mittelwert und ist im Streudiagramm als Punkt dargestellt. Wären beide Werte gleich, würden alle Punkte auf einer Geraden liegen, die die Nullpunkte beider Achsen schneidet. Für eine Person ergibt sich beispielsweise ein Faktorwert von 78 und ein Abweichungswert vom Mittelwert von 98. Das ist ein Unterschied von mehr als 1/3 Standardabweichungen (R: sd(dev_almd_mean); Ergebnis R: [1] 58.8058). Es handelt sich nicht einmal um ein extremes Beispiel, wie man an der Verteilung der Differenzen in Abbildung 4.30 sieht (R-Bildung der Differenz: diff_factor_dev_mean <- (f_almd - dev_almd_mean); R-Histogramm: hist(diff_factor_dev_mean)). Es kommen noch größere Abweichungen vor. Das heißt, im Einzelfall kann der Faktorwert unter Berücksichtigung korrelierter Fehlervariablen 0.6 Standardabweichungen vom Abweichungswert vom Mittelwert im TAP-Alertness-Test abweichen, selbst wenn beide Variablen 0.98 korrelieren. Einzelne Personen werden daher im Hinblick auf ihre Leistung durch die beiden unterschiedlichen Möglichkeiten, die Leistung der Person zu schätzen, unterschiedlich bewertet. Dies kann insbesondere bei Werten im Mittelbereich des Leistungsspektrums unter Berücksichtigung der Konfidenzintervalle zu anderen Einschätzungen der Leistung von einzelnen Personen führen (siehe ***Kapitel 9***).

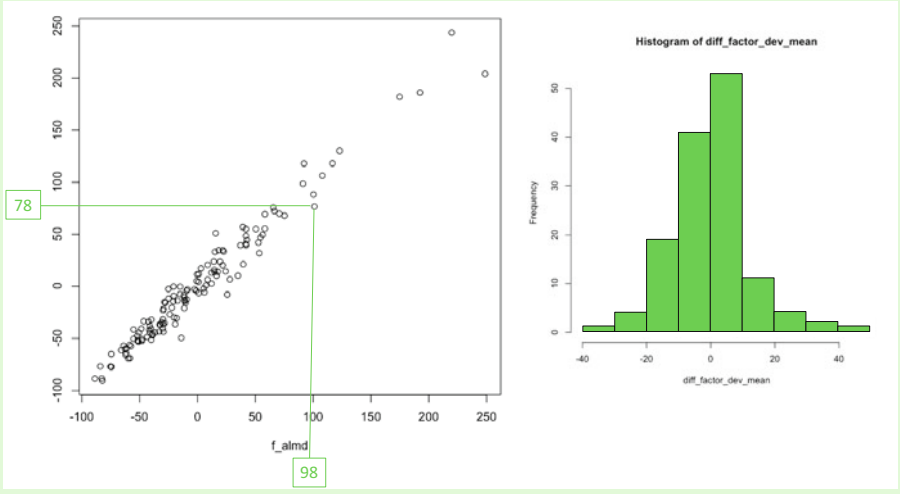



Abbildung 4.30: Korrelation zwischen unstandardisierten Faktorwerten eines essenziell tau-äquivalenten Modells und Abweichungswerten vom Mittelwert in der Alertness.

## Z U S A M M E N F A S S U N G

Nur unter bestimmten Voraussetzungen kann man mithilfe von Schätzfunktionen die Reliabilität von Itemmittelwerten schätzen. Die Voraussetzung dafür ist, dass die Items einer latenten Variablen (1) unkorrelierte Fehlervariablen aufweisen, (2) der wahre Wert eines Items für jede Person in einer genau definierten Beziehung zur latenten Variablen steht und berücksichtigt wird und klar ist, ob (3) alle Items gleich messgenau sind. Erst die zweite Voraussetzung führt dazu, dass die Modelle mithilfe von Strukturgleichungsmodellen prüfbar werden. Denn im Rahmen von Strukturgleichungsmodellen können die fünf hier dargestellten Modelle der klassischen Testtheorie geprüft werden. Müssen diese Modelle durch die empirischen Daten verworfen werden, müssen entweder mehrdimensionale Modelle verwendet werden oder man nimmt eine „Verschmutzung" des Itemmittelwerts der Items einer latenten Variable in Kauf. Im Folgenden sind die fünf besprochenen Testmodelle noch einmal im Kurzüberblick dargestellt.

**Parallele Items** messen **dasselbe Merkmal**, und nur das, mit **derselben Einheit** und zwar **für alle Items gleich genau**. Eine Schätzung der Reliabilität über den Mittelwert der Items einer latenten Variablen kann mithilfe der **Spearman-Brown-** oder **Cronbach-alpha-** oder **Omega**-Schätzfunktion erfolgen. Folgende Bedingungen müssen erfüllt sein:

(1) $\tau_i = \theta \rightarrow$ gleiche Einheit

(2) $\mathrm{VAR}(\varepsilon_i) = \mathrm{VAR}(\varepsilon_j) \rightarrow$ gleiche Messgenauigkeit der Items

(3) $\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0 \rightarrow$ unkorrelierte Fehlervariablen

**Essenziell parallele Items** messen **dasselbe Merkmal**, und nur das, mit **derselben Einheit**, und zwar **für alle Items gleich genau.** Die **wahren Werte der Items** sind lediglich um eine **additive Konstante von den Werten der latenten Variablen verschoben**. Eine Schätzung der Reliabilität über den Mittelwert der Items einer latenten Variablen kann mithilfe der **Spearman-Brown-** oder **Cronbach-alpha-** oder **Omega**-Schätzfunktion erfolgen. Folgende Bedingungen müssen erfüllt sein:

(1) $\tau = \theta + \sigma_i \rightarrow$ gleiche Einheit, um Konstante verschobener Wert

(2) $\mathrm{VAR}(\varepsilon_i) = \mathrm{VAR}(\varepsilon_j) \rightarrow$ gleiche Messgenauigkeit der Items

(3) $\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0 \rightarrow$ unkorrelierte Fehlervariablen

**Tau-äquivalente Items** messen **dasselbe Merkmal**, und nur das, mit **derselben Einheit**, und zwar unterschiedlich **genau für alle Items**. Eine Schätzung der Reliabilität über den Mittelwert der Items einer latenten Variablen kann mithilfe der **Cronbach-alpha-** oder **Omega**-Schätzfunktion erfolgen. Folgende Bedingungen müssen erfüllt sein:

(1) $\tau_i = \theta \rightarrow$ gleiche Einheit

(2) $\mathrm{VAR}(\varepsilon_i) \neq \mathrm{VAR}(\varepsilon_j) \rightarrow$ Messgenauigkeit der Items unterschiedlich

(3) $\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0 \rightarrow$ unkorrelierte Fehlervariablen

**Essenziell Parallele Items** messen **dasselbe Merkmal**, und nur das, mit **derselben Einheit** und unterschiedlich **genau für alle Items.** Die **wahren Werte der Items** sind lediglich um eine **additive Konstante von den Werten der latenten Variablen verschoben**. Eine Schätzung der Reliabilität über den Mittelwert der Items einer latenten Variablen kann mithilfe der **Cronbach-alpha-** oder **Omega**-Schätzfunktion erfolgen. Folgende Bedingungen müssen erfüllt sein:

(1) $\tau_i = \theta + \sigma_i \rightarrow$ gleiche Einheit, um Konstante verschobener Wert
(2) $\mathrm{VAR}(\varepsilon_i) \neq \mathrm{VAR}(\varepsilon_j) \rightarrow$ Messgenauigkeit der Items unterschiedlich
(3) $\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0 \rightarrow$ unkorrelierte Fehlervariablen

**Tau-kongenerische Items** messen **dasselbe Merkmal**, und nur das, mit **einer anderen Einheit**, und zwar unterschiedlich **genau für alle Items.** Die **wahren Werte der Items** sind sowohl um eine **additive** als auch durch eine **multiplikative Konstante von den Werten der latenten Variablen verschoben**. Eine Schätzung der Reliabilität über den Mittelwert der Items einer latenten Variablen kann mithilfe der **Omega**-Schätzfunktion erfolgen. Folgende Bedingungen müssen erfüllt sein:

(1) $\tau_i = \beta_i \cdot \theta + \sigma_i \rightarrow$ unterschiedliche Einheit
(2) $\mathrm{VAR}(\varepsilon_i) \neq \mathrm{VAR}(\varepsilon_j) \rightarrow$ gleiche Messgenauigkeit der Items
(3) $\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0 \rightarrow$ unkorrelierte Fehlervariablen

Aus diesen Annahmen ergeben sich für zufällige Personen überprüfbare Testmodelle der Klassischen Testtheorie (siehe Abbildung 4.31). Überprüft wird zum einen die Gleichheit der Erwartungswerte der Items ($\sigma_i = \sigma_j$, siehe Spalte Modell), die Gleichheit der Fehlervarianzen der Items ($\mathrm{VAR}(\varepsilon_i) = \mathrm{VAR}(\varepsilon_j)$), die Festlegung der Steigungsparameter auf den Wert Eins ($\beta_i = \beta_j = 1$, siehe Spalte Modell), die Annahme unkorrelierter Fehlervariablen ($\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0$).

| Äquivalenz | Modell | Bezug von $\theta$ zu $\tau_i$ | $\mathrm{COV}(\varepsilon_i, \varepsilon_j) = 0$ | $\mathrm{VAR}(\varepsilon_i) = \mathrm{VAR}(\varepsilon_j)$ |
|---|---|---|---|---|
| Grundmodell | $X_i = \tau_i + \varepsilon_i$ | | | |
| Parallel | $X_i = \theta + \varepsilon_i$ | $\tau_i = \theta$ | Ja | Ja |
| Essenziell parallel | $X_i = \theta + \sigma_i + \varepsilon_i$ | $\tau_i = \theta + \sigma_i$ | Ja | Ja |
| Tau-äquivalent | $X_i = \theta + \varepsilon_i$ | $\tau_i = \theta$ | Ja | Nein |
| Essenziell tau-äquivalent | $X_i = \theta + \sigma_i + \varepsilon_i$ | $\tau_i = \theta + \sigma_i$ | Ja | Nein |
| Tau-kongenerisch | $X_i = \theta \cdot \beta_i + \sigma_i + \varepsilon_i$ | $\tau_i = \beta_i \cdot \theta + \sigma_i$ | Ja | Nein |

Abbildung 4.31: Zusammenfassende Darstellung der Klassischen Testmodelle für zufällige Personen.

*Häufige Missverständnisse in der Praxis.* Zum Anschluss der Zusammenfassung sollen häufige Missverständnisse berichtet werden, die in der Praxis bezüglich der Evaluation eines psychometrischen Tests mithilfe der Klassischen Testtheorie und ihrer Modelle bestehen:

- Reliabilitäten werden geschätzt, obwohl die Items nicht auf unkorrelierte Fehlervariablen überprüft wurden und diese Prüfung erfolgreich war.
- Eine hohe Reliabilität (z. B. hohes Cronbach-alpha) wird mit statistischer Eindimensionalität gleichgesetzt: Die Reliabilität ist aber kein Index, der angibt, inwieweit Items eines Tests statistisch eindimensional sind, sondern setzt unkorrelierte Fehlervariablen als Modellannahme voraus.
- Die Klassischen Testmodelle werden bei der Angabe der Reliabilitäten nicht betrachtet: z. B. wird Cronbach-alpha auch für tau-kongenerische Messungen angegeben oder es werden dem Testanwender ohne Rückgriff auf die Klassischen Modelle mehrere Reliabilitätsschätzungen angeboten. Cronbach-alpha ist eine Mindestschätzung der Reliabilität im Rahmen des tau-kongenerischen Modells und kann damit immer als untere Schranke der Reliabilität interpretiert werden, wenn lokale Unabhängigkeit vorliegt. Daher hat Cronbach-alpha einen herausragenden Status unter

den Reliabilitätsschätzern, wenn Testinstrumente in ihrer Präzision verglichen werden sollen. Cronbach-alpha wird auch für die meisten Tests angegeben. Im Rahmen der psychometrischen Einzelfalldiagnostik ist es jedoch wünschenswert möglichst exakte Reliabilitätsschätzungen zu erhalten, um die Konfidenzintervalle so klein wie möglich zu halten.

- Zusätzliche Items erhöhen nicht immer die Reliabilität des Mittelwerts, sondern nur bei parallelen und essenziell parallelen Modellen.
- Split-Half, Paralleltest und Retest-Methoden sind für die Überprüfung unkorrelierter Fehlervariablen nicht geeignet, trotzdem werden die Methoden angewandt. Ein Test sollte mindestens vier Testteile aufweisen, damit eine Prüfung der klassischen Testmodelle möglich ist.
- Gleichlautende Items werden zur Verlängerung eines Tests hinzugefügt, um die Reliabilitätsschätzung eines Fragebogens zu erhöhen. Dies ist jedoch aus zwei Gründen extrem schädlich. Erstens weist der Test dann einen verzerrten Personenparameter auf: Einige gleichlautende Items bestimmen den Mittelwert über die Items einer latenten Variable durch ihr mehrfaches Vorkommen mehr als andere Items. Damit ergibt sich ein Bias bzw. eine Verzerrung des Mittelwerts. Zweitens zeigt die Praxis, dass diese gleichlautenden Items fast ausnahmslos korrelierte Fehler verursachen. Werden diese korrelierten Fehler bei der Reliabilitätsschätzung nicht berücksichtigt, verzerren sie die Reliabilitätsschätzung. Darüber hinaus liegt dann als weitere Folge statistische Eindimensionalität nicht mehr vor, es entstehen sogenannte lokale Abhängigkeiten.
- Immer komplexere Modelle werden an die Daten angepasst, ohne einen Rückgriff auf die Theorie. Modelle mit mehr zu schätzenden Parametern beschreiben die Daten in der Regel immer besser. Es werden also Modelle so lange an die Daten angepasst, bis ein Modell einen nicht signifikanten Hypothesentest aufweist oder die Fit-Indizes einigermaßen passen. Häufig wird dieses Vorgehen aus der Angst heraus gewählt, dass man sich bei einem „nicht passenden Modell" angreifbar macht. Allerdings kennt niemand das echte Populationsmodell. Daher kann man sich auch mithilfe einer theoretisch nachvollziehbaren Begründung für ein schlechter passendes Modell entscheiden gegenüber einem passenden oder an die Daten angepassten Modell.

**Z U S A M M E N F A S S U N G**

## 4.4.9 Itemtrennschärfe

*Ziel des Abschnitts.* In diesem Abschnitt wird die Trennschärfeanalyse als Teil einer Itemanalyse kurz eingeführt. Die Trennschärfeanalyse wird in der Praxis sehr häufig angewandt, daher ergibt sich die Notwendigkeit der Darstellung. Sie liefert in Abhängigkeit des angewandten Klassischen Testmodells nur wenig und nur sehr grobe Informationen über die Items einer latenten Variablen. Eine besondere Herausforderung stellen Trennschärfeanalysen mit diskreten Itemantworten dar. Zunächst wird definiert, was eine Trennschärfe ist und wie man sie interpretiert. Im Anschluss werden unterschiedliche Arten von Trennschärfen dargestellt und schließlich wird kurz auf Alternativen eingegangen.

**Trennschärfe**

Inhaltlich gibt eine Trennschärfe an, wie gut ein Item das angestrebte Konstrukt bzw. die angestrebte latente Variable misst. Sie wird mithilfe einer **Pearson-Korrelation** angegeben. Rechnerisch ist eine **Trennschärfe** nichts anderes als die Kor-

relation **eines Items** mit dem Mittelwert oder Summenwert der **restlichen** Items einer latenten Variablen. Die Mittelwerte bzw. Summenwerte der restlichen Items dienen dabei als eine Art Näherung für die Ausprägung der Personen auf der latenten Variablen.

*Trennschärfen.* Die Trennschärfe kann wie folgt am einfachsten berechnet werden:

$$r_{i(\overline{x}-i)} = \frac{\text{cov}_{i(\overline{x}-i)}}{s_i \cdot s_{\overline{x}-i}}$$

- $r_{i(\overline{x}-i)}$ = Korrelation des Items $i$ mit dem Mittelwert der restlichen Items einer latenten Variablen ohne das betrachtete Item $i$
- $s_i$ = Standardabweichung des Items $i$
- $s_{\overline{x}-i}$ = Standardabweichung des Mittelwerts der restlichen Items der latenten Variablen ohne das betrachtete Item $i$
- $\text{cov}_{i(\overline{x}-i)}$ = Kovarianz des Items $i$ mit dem Mittelwert der restlichen Items einer latenten Variablen ohne das betrachtete Item $i$

*Beurteilung von Trennschärfen.* Die Itemverteilungen können sich auf die Höhe der Trennschärfen auswirken. Unterscheidet sich die Schiefe eines Items von der Schiefe des korrigierten Skalenwerts, fällt die Trennschärfe geringer aus als bei gleicher Schiefe. Es ist jedoch in vielen Fällen, außer bei Schnelligkeitstests mit nahezu identischen Items, notwendig, dass ein Test Items mit unterschiedlicher Schiefe und damit unterschiedlichen Itemparametern enthält. Erst durch Items mit unterschiedlicher Schiefe bzw. Trennschärfeparametern können wir zwischen Personen verschiedener Fähigkeits- und Eigenschaftsausprägung unterscheiden. Daher ist bei einer späteren Itemselektion darauf zu achten, dass die Items nicht aufgrund ihrer Trennschärfe aus dem Test entfernt werden, sondern auch die Schwierigkeit, die Standardabweichung und die Verteilung der Items mitberücksichtigt werden. Wird ein Item aus bestimmten Gründen aus dem Test entfernt, muss kritisch geprüft werden, ob sich dadurch der Messgegenstand des Tests verändert, weil nicht mehr das vorher definierte Konstrukt erfasst wird. Will man die Konstruktdefinition nicht verändern, kann eine Umformulierung eines Items hilfreich sein sowie eine erneute Evaluation des Items.

Bei der Beurteilung der Trennschärfe ist wichtig, dass die Unterscheidungen, die anhand des Items getroffen werden, beispielsweise Item gelöst oder nicht gelöst, im Sinne der zu messenden Fähigkeit oder Eigenschaft ausfallen. Das heißt, dass Personen, die ein Item lösen, auch einen hohen Mittelwert der restlichen Items erhalten und Personen, die ein Item nicht lösen, einen niedrigen Mittelwert der restlichen Items. Diese Erwartung muss nicht eintreffen, denn einzelne Items können auch entgegen der Erwartung mit dem Mittelwert der restlichen Items korrelieren. In diesem Fall ergeben sich negative Trennschärfen. Items mit negativen Trennschärfen sind in der Regel für die Testkonstruktion ungeeignet. Werden negativ gepolte Items (positiv formuliert „Ich bin glücklich." vs. negativ formuliert „Ich bin traurig."), wie man sie mitunter in Fragebogen findet, verwendet, sollten diese vor der Trennschärfeanalyse umgepolt werden. Items mit Nullkorrelationen sind ebenfalls für die Testkonstruktion ungeeignet, denn sie zeigen, dass das Item nicht mit der zu messenden latenten Variablen zusammenhängt und etwas anderes misst. Zur Testkonstruktion sollten große Stichproben (mindestens $n > 250$, besser $n = 400$ oder größer) verwendet werden, um einigermaßen präzise Schätzungen der Trennschärfen zu erhalten. Trennschärfen, die in einem

solchen Fall statistisch nicht signifikant sind, sollten aus dem Test entfernt werden. Die Höhe der Trennschärfe ist kein Kriterium für die Itemselektion. Folgende Faustregeln können für die Trennschärfeanalyse verwendet werden:

- Im parallelen und essenziell parallelen Modell müssen die Trennschärfen gleich hoch sein.
- Im tau-kongenerischen Modell dürfen Trennschärfen in ihrer Höhe variieren. Im tau-kongenerischen Modell sollten die Trennschärfen zumindest statistisch signifikant sein. Sie können aus dem Test entfernt werden, wenn man glaubt, dass die niedrige Trennschärfe nicht an der Itemformulierung liegt. Vermutet man, dass die niedrige Trennschärfe an der Itemformulierung liegt, sollte diese geändert werden, um eine hohe Inhaltsvalidität zu sichern.
- In allen Fällen muss vorher mithilfe von Strukturgleichungsmodellen (siehe ***Kapitel 8***) überprüft werden, ob unkorrelierte Fehlervariablen vorliegen.

### Arten von Trennschärfekoeffizienten



*Part-Whole-Korrektur.* In den meisten Fällen wird nicht der Aufwand betrieben, alle Mittelwerte von Items einer latenten Variablen zu bilden, bei der jeweils ein Item nicht berücksichtigt wird, um dann im Anschluss die Korrelation eines Items mit dem Mittelwert der restlichen Items einer latenten Variablen zu bilden. In der Regel wird die Korrelation zwischen dem Mittelwert aller Items einer latenten Variablen und dem entsprechenden Item betrachtet und die Korrelation statistisch korrigiert bzw. bereinigt. Diese Korrektur wird als **Part-Whole-Korrektur** oder **Teil-Ganzes-Korrektur** bezeichnet. Würde eine solche Korrektur nicht durchgeführt werden, würde das Item auch mit sich selbst in der Gesamtskala korrelieren. Diese Eigenkorrelation würde zu einer Überschätzung der Trennschärfe führen. Je mehr Items ein Test enthält, desto geringer fallen die Auswirkungen der statistischen Korrektur auf die Trennschärfe aus. Der Grund dafür liegt darin, dass mit zunehmender Itemanzahl der Beitrag eines einzelnen Items zum Mittelwert oder Summenwert geringer ausfällt. Eine positive Korrelation zwischen dem Item und dem Mittelwert der restlichen Items einer latenten Variablen bedeutet, je höher die Reaktionszeit auf ein Item ausfällt, desto höher fällt auch die durchschnittliche mittlere Reaktionszeit auf den restlichen Items aus.

*Selektionskennwert (SK).* Um die Problematik mit den Itemverteilungen etwas abzumildern, wurde ein sogenannter Selektionskennwert entwickelt. Er basiert auf der Trennschärfe. Die Verwendung des Selektionskennwerts führt dazu, dass die Trennschärfe von Items mit extremer Schwierigkeit nach oben korrigiert wird. Er ist insbesondere für dichotome Itemantworten und Itemantworten mit mehr als zwei geordneten Antwortkategorien gedacht. Der Selektionskennwert kann wie folgt bestimmt werden (Lienert & Raatz, 1998):

$$SK = \frac{r_{i(\bar{x}-i)}}{2 \cdot s_i}$$

- $r_{i(\bar{x}-i)}$ = Korrelation des Items $i$ mit dem Mittelwert der restlichen Items einer latenten Variablen ohne das betrachtete Item $i$
- $s_i$ = Standardabweichung des Items $i$

Betrachten wir zur Veranschaulichung folgendes Beispiel: Nehmen wir an, dass der Itemparameter für ein beliebiges dichotomes Item $i$ mit $p_i = 0.80$ sehr hoch ausfällt. Der Wert bedeutet, dass 80 Prozent der Stichprobe dieses Item richtig gelöst haben. Nehmen wir weiter an, es ergibt sich für Item $i$ auch eine geringe Trennschärfe von

$r_{i(\overline{x}-i)} = 0.10$. Die Standardabweichung für dichotome Items ergibt sich direkt aus dem Itemparameter: $s_i = \sqrt{p_i \cdot (1 - p_i)} = \sqrt{0.80 \cdot 0.20} = 0.40$. Es ergibt sich eine Standardabweichung von 0.40 für Item *i*. Betrachten wir nun ein Item *j* mit einem Itemparameter von $p_j = 0.50$ und einer Trennschärfe von $r_{j(\overline{x}-j)} = 0.50$. Es ergibt sich eine Standardabweichung von $s_j = \sqrt{p_j \cdot (1 - p_j)} = \sqrt{0.50 \cdot 0.50} = 0.50$. Berechnet man nun für Item *j* den Selektionskennwert, muss die Trennschärfe $r_{j(\overline{x}-j)} = 0.50$ durch die mit der Zahl zwei multiplizierte Standardabweichung ($s_i = 0.50$) geteilt werden. Der Selektionskennwert beträgt demnach $SK = 0.50$ und entspricht exakt der Trennschärfe. Für das Item *i* mit der höheren psychometrischen Schwierigkeit wird die Trennschärfe von $r_{i(\overline{x}-i)} = 0.10$ zweimal durch die Standardabweichung 0.80 ($2 \cdot s_i = 2 \cdot 0.40 = 0.80$) geteilt. Somit fällt der Selektionskennwert ($SK = \frac{0.10}{0.80} = 0.125$) höher als die Trennschärfe von 0.10 aus. Je weiter sich der Schwierigkeitsindex den Randbereichen von null und von eins annähert, desto stärker fällt die Korrektur aus. Damit werden Items mit extremeren Schwierigkeitsindizes im Rahmen der Testanalyse nicht so leicht aus einem Test entfernt. Items mit einem Schwierigkeitsindex von eins oder null können mit dieser Formel nicht korrigiert werden. Eine Division durch zweimal null ist nicht erlaubt. Solche Items sind für einen Test auch ungeeignet, denn sie korrelieren nicht mit der latenten Variablen und messen etwas anderes. Ist die Trennschärfe eines Items eins, benötigt man entweder nur ein Item oder das Item ist redundant bzw. überflüssig. Spezifische Möglichkeiten der Itemselektion beschreiben Lienert und Raatz (1998) oder Fisseni (1997, S. 60 und S. 62). Bühner et al. (2006) haben für den I-S-T 2000 R im Grundmodul gezeigt, dass die Verwendung des Selektionskennwerts immer noch dazu führt, dass viele Items geringe Trennschärfen aufweisen. Eine Analyse mithilfe von Testmodellen für diskrete Itemantworten ist hier vorteilhafter (z. B. Rasch-Modell), da anhand einer Trennschärfeanalyse deutlich weniger Items ausgeschlossen gewesen wären als nötig.

*Minderungskorrigierte Trennschärfen.* In R im Paket `psych` (Revelle, 2017) wurden weitere Trennschärfekoeffizienten vorgeschlagen. Bei dieser Trennschärfe wird eine Minderungskorrektur durchgeführt. Die Mittelwerte der restlichen Items auf der latenten Variablen unterscheiden sich in Abhängigkeit der Itemzusammensetzung. Eine nicht perfekte Reliabilität dieser Mittelwerte mindert die Korrelationen der Mittelwerte mit den Items. Die Korrektur wird als einfache Minderungskorrektur bezeichnet und korrigiert die Trennschärfen nach oben. Sie stellt damit vor allem in Rechnung, dass sich die Mittelwerte der restlichen Items einer latenten Variablen in der Reliabilität unterscheiden. Nehmen wir an, die Part-Whole-korrigierte Trennschärfe eines Items beträgt 0.72 und die Reliabilität des Mittelwerts der restlichen Items der latenten Variablen beträgt 0.89, dann ergibt sich folgender Schätzwert für die einfach minderungs- und Part-Whole-korrigierte Trennschärfe:

$$r_{i(\overline{x}-1),korr} = \frac{r_{i(\overline{x}-i)}}{\sqrt{\text{rel}_{\overline{x}-i}}} = \frac{.718}{\sqrt{.89}} = \frac{.718}{.943} = .761$$

- $r_{i(\overline{x}-1),korr}$ = einfach minderungskorrigierte Korrelation des Items *i* mit dem Mittelwert der restlichen Items einer latenten Variablen ohne das betrachtete Item *i*
- $r_{i(\overline{x}-i)}$ = Korrelation des Items *i* mit dem Mittelwert der restlichen Items einer latenten Variablen ohne das betrachtet Item *i*
- $\text{rel}_{\overline{x}-i}$ = Schätzwert der Reliabilität des Mittelwertes der Items der latenten Variablen ohne das betrachtete Item *i*

## Experten 4.11 Doppelte und einfache Minderungskorrektur

Wie kommt man auf die Formel für die Minderungskorrektur? Dies wird im Folgenden kurz erläutert: Wir gehen zunächst davon aus, dass man die Itemantworten auf beiden Items wie folgt schreiben kann: $X_j = \tau_j + \varepsilon_j$ und $X_i = \tau_i + \varepsilon_i$. Für beide Items setzen sich die Itemantwortvariablen aus der True-Score-Variablen $(\tau_j, \tau_i)$ plus den Fehlervariablen $(\varepsilon_i, \varepsilon_i)$ zusammen. Wir haben auch gesehen, dass wir im Falle paralleler Messungen die True-Score-Variablen $(\tau_j, \tau_i)$ durch die latenten Variablen $(\theta_j, \theta_i)$ ersetzen können. Wir möchten in diesem Fall die „wahre" True-Score-Korrelation oder latente Variablen-Korrelation ermitteln.

Betrachten wir zunächst folgende Gleichungen:

$$X_j = \theta_j + \varepsilon_j \quad \text{und} \quad X_i = \theta_i + \varepsilon_i$$

Wir gehen davon aus, dass die Fehlervariablen unkorreliert sind ($\text{COR}(\varepsilon_i, \varepsilon_j) = 0$) und darüber hinaus die Fehlervariablen auch nicht mit den True-Score-Variablen korrelieren: $\text{COR}(\varepsilon_i, \tau_j) = 0$ und $\text{COR}(\varepsilon_j, \tau_i)$. Wenn wir für diesen Fall die Kovarianzen der beiden Itemantwortvariablen betrachten, erhalten wir:

$$\begin{aligned}
\text{COV}\left(X_i, X_j\right) &= \text{COV}\left(\theta_i + \varepsilon_i, \theta_j + \varepsilon_j\right) \\
\text{COV}\left(\theta_i + \varepsilon_i, \theta_j + \varepsilon_j\right) &= \text{COV}\left(\theta_i, \theta_j\right) + \text{COV}\left(\varepsilon_i, \theta_j\right) + \text{COV}\left(\varepsilon_j, \theta_i\right) + \text{COV}\left(\varepsilon_i, \varepsilon_j\right) \\
\text{COV}\left(\theta_i + \varepsilon_i, \theta_j + \varepsilon_j\right) &= \text{COV}\left(\theta_i, \theta_j\right) + 0 + 0 + 0 = \text{COV}\left(\theta_i, \theta_j\right)
\end{aligned}$$

In einem nächsten Schritt betrachten wir die Definition der Reliabilität:

$$\text{REL}\left(X_i\right) = \frac{\text{VAR}\left(\tau_i\right)}{\text{VAR}\left(X_i\right)} \text{ bzw. } \text{REL}\left(X_i\right) = \frac{\text{VAR}\left(\theta_i\right)}{\text{VAR}\left(X_i\right)}$$

Wir können diese Formel umstellen, indem wir sie nach $\text{VAR}(X_i)$ auflösen:

$$\text{VAR}\left(X_i\right) \cdot \text{REL}\left(X_i\right) = \text{VAR}\left(\theta_i\right) \text{ bzw. } \text{VAR}\left(X_j\right) \cdot \text{REL}\left(X_j\right) = \text{VAR}\left(\theta_j\right)$$

Setzen wir nun diese Formeln ($\text{VAR}(X_i) \cdot \text{REL}(X_i)$ bzw. $\text{VAR}(X_j) \cdot \text{REL}(X_j)$) für die Varianz der wahrten Werte ($\text{VAR}(\theta_i)$ und $\text{VAR}(\theta_j)$) in die Formel für die Korrelation zwischen zwei latenten Variablen ein:

$$\text{COR}\left(\theta_i, \theta_j\right) = \frac{\text{COV}\left(\theta_i, \theta_j\right)}{\sqrt{\text{VAR}\left(\theta_i\right)} \cdot \sqrt{\text{VAR}\left(\theta_j\right)}}$$

Wir erhalten dann folgende Formel für die Korrelation:

$$\text{COR}\left(\theta_i, \theta_j\right) = \frac{\text{COV}\left(X_i, X_j\right)}{\sqrt{\text{VAR}\left(X_i\right) \cdot \text{REL}\left(X_i\right)} \cdot \sqrt{\text{VAR}\left(X_j\right) \cdot \text{REL}\left(X_j\right)}}$$

Wir können diese Formel weiter umstellen:

$$\text{COR}\left(\theta_i, \theta_j\right) = \frac{\text{COV}\left(X_i, X_j\right)}{\frac{\sqrt{\text{VAR}(X_i)} \cdot \sqrt{\text{VAR}(X_j)}}{\sqrt{\text{REL}(X_i)} \cdot \sqrt{\text{REL}(X_j)}}}$$

Daraus ergibt sich dann die **doppelte Minderungskorrektur**:

$$\mathrm{COR}(\theta_i, \theta_j) = \frac{\mathrm{COR}(X_i, X_j)}{\sqrt{\mathrm{REL}(X_i)} \cdot \sqrt{\mathrm{REL}(X_j)}}, \quad \text{da } \mathrm{COR}(X_i, X_j) = \frac{\mathrm{COV}(X_i, X_j)}{\sqrt{\mathrm{VAR}(X_i)} \cdot \sqrt{\mathrm{VAR}(X_j)}}$$

Wird nur in einer Variablen für die nicht perfekte Reliabilität kontrolliert, vereinfacht sich die Formel weiter zu:

$$\mathrm{COR}(\theta_i, X_j) = \frac{\mathrm{COR}(X_i, X_j)}{\sqrt{\mathrm{REL}(X_i)}}$$

Diese Formel wird als **einfache Minderungskorrektur** bezeichnet.

### *Alternativen zu Trennschärfeanalysen*

*Faktorenanalyse oder Trennschärfeanalyse?* Trennschärfeanalysen berücksichtigen nur die Korrelation jeweils eines Items mit einem groben Schätzwert für die Ausprägung auf einer einzigen latenten Variablen. Wir finden in unserem Beispiel nur heraus, ob ein Item hoch mit dem Mittelwert aus den restlichen Items korreliert, nicht aber, ob es beispielsweise auch mit den Werten anderer latenter Variablen korreliert. Diese Information wäre aber wünschenswert.

Trennschärfeanalysen werden immer noch sehr häufig durchgeführt, sind aber nicht unproblematisch. Zum einen ist der Mittelwert der restlichen Items nicht immer ein guter Schätzwert für das Konstrukt bzw. die latente Variable. Zum anderen wird die Itemtrennschärfe häufig bei Items mit diskreten Itemantworten angewandt, was teilweise zu einer massiven Unterschätzung der Itemtrennschärfen führen kann. Daher ist von der Durchführung einer Trennschärfeanalyse eher abzuraten und entsprechende faktorenanalytische Modelle mit den entsprechenden Schätzmethoden (z. B. WLSMV) sind zu bevorzugen oder man sollte gleich entsprechende Testmodelle für diskrete Itemantworten durchführen (siehe ***Kapitel 5***).

## 4.5 Schwierigkeits-, Trennschärfe- und Reliabilitätsanalyse mit SPSS und R

*Ziel des Abschnitts*. In diesem Abschnitt werden Schwierigkeits-, Trennschärfe- und Reliabilitätsanalysen in SPSS und R dargestellt. Diese Analysen finden sich als Standard in vielen Testmanualen. Bis auf die hier dargestellte Reliabilitätsanalyse können und sollten die hier aufgeführten Analysen durch Analysen mithilfe von Strukturgleichungsmodellen ersetzt werden. Im Folgenden wird zunächst die Schwierigkeitsanalyse in SPSS und R dargestellt, im Schluss die Trennschärfeanalyse und schließlich die Reliabilitätsanalyse.

Um die Inhalte der folgenden Abschnitte besser zu verstehen, sind vor allem die Kenntnisse des Abschnitts 2.2 (S. 38) in Bühner & Ziegler (2017) günstig. Dort werden deskriptive Statistiken wie Mittelwerte, Standardabweichungen, Korrelationen und Schiefe sowie grafische Darstellungen von Histogrammen, Boxplots und Streudiagrammen beschrieben.

# Copyright