www.pearson-studium.de
CWS
COMPANION WEBSITE

ps
psychologie

Markus Bühner

Einführung in die Test- und Fragebogenkonstruktion

3., aktualisierte Auflage

PEARSON
Studium

Markus Bühner

Einführung in die Test- und Fragebogenkonstruktion

3., aktualisierte und erweiterte Auflage

PEARSON
Studium

ein Imprint von Pearson Education
München • Boston • San Francisco • Harlow, England
Don Mills, Ontario • Sydney • Mexico City
Madrid • Amsterdam

wenn kein spezifischer Wert im Testhandbuch angegeben wird, sollte eine andere geeignete Reliabilitätsschätzung herangezogen werden.

Die Wahl der geeigneten Reliabilitätsschätzung hängt von der Fragestellung ab. Soll eine **Prognose** (Längsschnittdiagnose) über die zukünftige Leistung gemacht werden, so sollte die **Retest-Korrelation** verwendet werden, da sie neben der Reliabilität auch ein Maß für die Stabilität eines Merkmals ist. Wenn keine Retest-Korrelation vorhanden ist, dann sollte die interne Konsistenz, wie Cronbach-α, die Split-Half-Koeffizienten oder am besten die Paralleltest-Reliabilität als Reliabilitätsschätzung herangezogen werden. Interessiert lediglich die Prüfung eines **aktuellen Status** (Querschnittdiagnose), sollte die **interne Konsistenz** bzw. die Split-Half- oder die Paralleltest-Korrelation als Reliabilität verwendet werden.

Im diagnostischen Prozess können verschiedene Fragestellungen interessant sein, die sich mithilfe der Regressions- bzw. Äquivalenzhypothese beantworten lassen. Zwei der möglichen Fragestellungen werden im Folgenden dargestellt und anhand von Beispielen erläutert.

ZUSAMMENFASSUNG

Folgende Checkliste soll noch einmal verdeutlichen, welche Schritte bei der Wahl der Einflussgrößen auf das Vertrauensintervall zu beachten sind:

- Fragestellung:
  - Einseitig: Es liegt eine begründete Vermutung vor, dass das zu erwartende Ergebnis in einen bestimmten Merkmalsbereich hineinfällt.
  - Zweiseitig: Es liegt **keine** Vermutung vor, in welchem Bereich das Ergebnis zu erwarten ist.
- Art des Vertrauensintervalls:
  - Äquivalenzhypothese kann in fast allen Fällen gewählt werden.
  - Regressionshypothese: Bei extremen Merkmalsausprägungen kann eine Korrektur des beobachteten Werts zur Mitte hin sinnvoll sein.
- Wahl der Sicherheitswahrscheinlichkeit:
  - Es ist günstig, sich zuerst an den negativen Konsequenzen für die Person zu orientieren und dann an der Messgenauigkeit des Messwerts. Ist der Test wenig messgenau, so dass Vertrauensintervalle resultieren, die sich vom unter- bis in den überdurchschnittlichen Bereich erstrecken, muss die Sicherheitswahrscheinlichkeit reduziert werden.
- Wahl der Reliabilität:
  - Querschnittdiagnose: Es ist relevant, wie eine Person am heutigen Tag abschneidet, eine Prognose auf zukünftiges Verhalten wird nicht gefordert. In diesem Fall kann die innere/interne Konsistenz verwendet werden.
  - Längsschnittdiagnose: Es wird gefordert, dass die Einschätzung über die Zeit hinweg stabil ist, beispielsweise im Rahmen der Eignungsdiagnostik. In diesem Fall kann die Retest-Korrelation als Reliabilitätsschätzung verwendet werden.

Es muss immer die Reliabilitätsschätzung aus der Normstichprobe verwendet werden, die der getesteten Person zugeordnet werden kann. Liegen verschiedene Reliabilitätsschätzungen vor, sollte die verwendet werden, für die die Voraussetzungen am ehesten erfüllt sind.

ZUSAMMENFASSUNG

Bei den nun folgenden Rechenbeispielen können Normwertgrenzen für Vertrauensintervalle auftreten, die nicht beobachtet werden können, beispielsweise eine untere Grenze von IQ = 102.2 und eine obere Grenze von 111.4. In diesem Fall beinhaltet das Vertrauensintervall Werte von 103 bis 111, denn ein Wert von 103 liegt noch innerhalb der angegebenen Grenzen des Vertrauensintervalls, ebenso ein Wert von 111. Werte für die untere und obere Grenze von 102 oder 112 liegen dagegen außerhalb der angegebenen Grenzen.

4.8.1 Vertrauensintervalle um den beobachteten Wert einer individuellen Testleistung

Wie gut kann ich mich auf das Testergebnis verlassen?

Wie wir aus *Kapitel 2* wissen, ist die Annahme, man könne Personen unter identischen Bedingungen wiederholt untersuchen, nicht haltbar. Wir wissen auch, dass sich Gedächtnis- und Übungseffekte (Fischer, 1974) auf die Testleistung auswirken können. Darüber hinaus ist eine Wiederholungsmessung nicht unendlich oft möglich. In seltenen Fällen bleibt die Zeit, eine Person zweimal zu testen. Daher besteht die Notwendigkeit, den Messfehler im Rahmen einer Messung zu ermitteln. Dazu stehen prinzipiell zwei Wege zur Verfügung: Man kann den Messfehler einer Messung mithilfe des Standardmessfehlers bestimmen oder mithilfe des Standardschätzfehlers. Die Basis dafür liefern zwei Hypothesen: Die Äquivalenz- und die Regressionshypothese.

4

Äquivalenzhypothese (Standardmessfehler)

Unter der Äquivalenzhypothese wird angenommen, dass der beobachtete Wert eine gute Schätzung des wahren Werts darstellt $(x_v = \hat{\tau}_v)$. Es wird weiter angenommen, dass der **Messfehler bei der Messung für jede Person gleich groß** ausfällt und folglich zwischen Personen mit unterschiedlicher Fähigkeits- oder Eigenschaftsausprägung nicht variiert. Um den Fehler zu bestimmen, wird nun zunächst als Vorstufe der Anteil der Unterschiedlichkeit zwischen den Personen geschätzt, der auf zufällige Störungen der Messung zurückzuführen ist. Es handelt sich dabei um den **nicht** reliablen Anteil der Merkmalsvarianz. Dieser Anteil wird mit folgender Formel bestimmt:

$$\sigma^2_{\mathrm{E}_x} = \sigma^2_x \cdot (1 - \rho_{tt})$$

- $\sigma^2_{\mathrm{E}_x}$ = quadrierter Standardmessfehler
- σ^2_x = Varianz des Messwerts
- ρ_{tt} = Reliabilität des Messwerts

Die Wurzel aus diesem unreliablen Anteil der Merkmalsvarianz stellt die so genannte Fehlerstreuung dar und wird als **Standardmessfehler** bezeichnet. Sie wird mithilfe der folgenden Formel ermittelt:

$$\sigma_{\mathrm{E}_x} = \sigma_x \cdot \sqrt{1 - \rho_{tt}}$$

Mithilfe des beobachteten Werts und des Standardmessfehlers sowie der Sicherheitswahrscheinlichkeit ergibt sich für das **Vertrauensintervall** laut Äquivalenzhypothese dann die folgende Formel für eine zweiseitige Fragestellung:

$$VI_{u,o} = x_v \pm \sigma_{E_x} \cdot z_{1-\frac{\alpha}{2}} = x_v \pm \sigma_x \cdot \sqrt{(1-\rho_{tt})} \cdot z_{1-\frac{\alpha}{2}}$$

Für die Berechnung eines Vertrauensintervalls bei einer einseitigen Fragestellung wird die folgende Formel verwendet:

$$VI_u = x_v - \sigma_{E_x} \cdot z_{1-\alpha} = x_v - \sigma_x \cdot \sqrt{(1-\rho_{tt})} \cdot z_{1-\alpha}$$

oder

$$VI_o = x_v + \sigma_{E_x} \cdot z_{1-\alpha} = x_v + \sigma_x \cdot \sqrt{(1-\rho_{tt})} \cdot z_{1-\alpha}$$

- $VI_{(u,o)}$ = Vertrauensintervall (untere, obere Grenze)
- VI_u = Vertrauensintervall untere Grenze
- VI_o = Vertrauensintervall obere Grenze
- x_v = beobachteter Normwert der Person v
- σ_{E_x} = Standardmessfehler
- σ_x = Standardabweichung des Messwerts
- ρ_{tt} = Reliabilität des Messwerts
- $z_{1-\alpha}$ = kritischer z-Wert der Standardnormalverteilung für eine bestimmte Irrtumswahrscheinlichkeit α bei einseitiger Testung
- $z_{1-\alpha/2}$= kritischer z-Wert der Standardnormalverteilung für eine bestimmte Irrtumswahrscheinlichkeit α bei zweiseitiger Testung

Regressionshypothese (Standardschätzfehler):

Anders als bei der Äquivalenzhypothese wird unter der Regressionshypothese angenommen, dass der wahre Wert einer Person erst aus dem beobachteten Wert geschätzt werden muss. Diese Schätzung erfolgt mithilfe des beobachteten Werts, der Reliabilität und dem Mittelwert der Rohwerte bzw. der verwendeten Normwerte. Kennt man die Verteilung der beobachteten Werte der Personen einer Population, ist der wahre Wert der Population der Erwartungswert der Verteilung der beobachteten Werte: E (T | X). Geht man von einer Normalverteilung der beobachteten Messwerte in der Population aus, weiß man, dass Personen mit extremen Messwerten nur selten beobachtet werden. Besitzt der Messwert eine Reliabilität von eins, stellt dies kein Problem dar. In diesem Fall entspricht der beobachtete Wert dem wahren Wert. Ist der Messwert jedoch nicht perfekt messgenau, variieren die beobachteten Werte einer Person um ihren konstanten wahren Wert. Wird nun ein extremer, nicht messgenauer Wert einer Person beobachtet, ist es wahrscheinlicher, dass der wahre Wert dieser Person vom beobachteten Wert in Richtung des wahren Werts der Population, also zur Mitte hin, abweicht. Denn der wahre Wert der Population ist der wahrscheinlichste Wert für eine Person, die aus dieser Population gezogen wurde. Es wird daher eine Korrektur hin zum wahren Wert der Population vorgenommen. Diese Korrektur fällt umso stärker aus, je weniger reliabel der Messwert gemessen werden kann. Der **wahre Wert** wird mithilfe der folgenden Formel geschätzt:

$$\hat{\tau}_v = \rho_{tt} \cdot x_v + \mu_x \cdot (1 - \rho_{tt})$$

- $\hat{\tau}_v$ = geschätzter wahrer Wert der Person v
- x_v = beobachteter Normwert der Person v
- μ_x = Mittelwert der Norm
- ρ_{tt}= Reliabilität des Messwerts

Der Standardschätzfehler charakterisiert die Vorhersagegenauigkeit des geschätzten wahren Werts. Der Fehler bei der Schätzung kann mithilfe der folgenden Formel ermittelt werden:

$$\sigma_{E_\tau} = \sigma_x \cdot \sqrt{\rho_{tt} \cdot (1 - \rho_{tt})}$$

- σ_{E_τ} = Standardschätzfehler
- σ_x = Standardabweichung des Messwerts
- ρ_{tt} = Reliabilität des Messwerts

Exkurs 4.7

Im Folgenden soll die Formel für den Standardschätzfehler weiter veranschaulicht werden: Betrachten wir dazu zunächst die Varianz der wahren Werte σ^2_T der Personen, die einen bestimmten Test bearbeitet haben. Wie wir wissen, ist die Varianz der wahren Werte wie folgt definiert:

$$\sigma^2_T = \rho_{tt} \cdot \sigma^2_x$$

Nun wird der **Fehler bei der Schätzung** (Schätzfehler**) bzw. Vorhersage der wahren Werte** bestimmt. Dazu wird wieder der nicht reliable Anteil des Messwerts $(1 - \rho_{tt})$ herangezogen:

$$\sigma^2_{E_\tau} = \sigma^2_T \cdot (1 - \rho_{tt}) = \rho_{tt} \cdot \sigma^2_x \cdot (1 - \rho_{tt})$$

Setzt man für σ^2_T wieder $\rho_{tt} \cdot \sigma^2_x$ in die Formel ein, ergibt sich $\rho_{tt} \cdot \sigma^2_x \cdot (1 - \rho_{tt})$.

Zieht man aus der gerade präsentierten Formel die Wurzel, erhält man den Standardschätzfehler:

$$\sigma_{E_\tau} = \sigma_x \cdot \sqrt{\rho_{tt} \cdot (1 - \rho_{tt})}$$

- σ_{E_τ} = Standardschätzfehler
- σ^2_T = Varianz der wahren Werte
- σ_x = Standardabweichung des Messwerts
- ρ_{tt}= Reliabilität des Messwerts

Nun kann das **Vertrauensintervall** um den geschätzten wahren Wert mithilfe des Standardschätzfehlers und der Sicherheitswahrscheinlichkeit für eine zweiseitige Fragestellung ermittelt werden:

$$VI_{u,o} = \hat{\tau}_v \pm \sigma_{E_\tau} \cdot z_{1-\frac{\alpha}{2}} = \hat{\tau}_v \pm \sigma_x \cdot \sqrt{\rho_{tt} \cdot (1-\rho_{tt})} \cdot z_{1-\frac{\alpha}{2}}$$

Das **Vertrauensintervall** um den geschätzten wahren Wert mithilfe des Standardschätzfehlers und der Sicherheitswahrscheinlichkeit für eine einseitige Fragestellung lautet:

$$VI_u = \hat{\tau}_v - \sigma_{E_\tau} \cdot z_{1-\alpha} = \hat{\tau}_v - \sigma_x \cdot \sqrt{\rho_{tt} \cdot (1-\rho_{tt})} \cdot z_{1-\alpha}$$

oder

$$VI_o = \hat{\tau}_v + \sigma_{E_\tau} \cdot z_{1-\alpha} = \hat{\tau}_v + \sigma_x \cdot \sqrt{\rho_{tt} \cdot (1-\rho_{tt})} \cdot z_{1-\alpha}$$

- $VI_{(u,o)}$= Vertrauensintervall (untere, obere Grenze)
- VI_u = Vertrauensintervall untere Grenze
- VI_o= Vertrauensintervall obere Grenze
- $\hat{\tau}_v$ = geschätzter wahrer Wert des Probanden
- σ_x = Standardabweichung des Messwerts
- σ_{E_τ} =Standardschätzfehler
- ρ_{tt}= Reliabilität des Tests
- $z_{1-\alpha}$= kritischer z-Wert der Standardnormalverteilung für eine bestimmte Irrtumswahrscheinlichkeit α bei einseitiger Testung
- $z_{1-\alpha/2}$= kritischer z-Wert der Standardnormalverteilung für eine bestimmte Irrtumswahrscheinlichkeit α bei zweiseitiger Testung

Voraussetzungen für die Anwendung von Standardmess- und Standardschätzfehler

Standardmessfehler Folgende Voraussetzungen müssen bei der Verwendung des Standardmessfehlers auf Basis der Äquivalenzhypothese beachtet werden: Die Fehlervarianz eines Tests ist in allen Skalenbereichen gleich groß (Homoskedastizität, vgl. Stelzl, 1993, S. 47, S. 55), die Messfehler und wahren Werte müssen normalverteilt sein. Dabei muss die Varianz der wahren Werte größer als null sein. Weiterhin darf sich auch innerhalb einer Normgruppe der Messfehler über Untergruppen nur geringfügig unterscheiden (Huber, 1973, S. 61). Das heißt, wird die Normstichprobe von 18- bis 25-jährigen männlichen Personen mit Abitur gebildet, darf der Messfehler innerhalb der Gruppe 18- bis 20-jähriger Personen mit Abitur nicht größer sein als innerhalb der Gruppe 21- bis 25-jähriger Personen mit Abitur. Da weder die wahren Werte noch die Messfehler für jede Person bekannt sind, ist eine Prüfung schwierig. Die einzige Möglichkeit der Prüfung besteht darin, das Histogramm des Testkennwerts zu betrachten und zu bewerten, ob für diesen Testwert eine Abweichung von einer Normalverteilung vorliegt.

Standardschätzfehler Bei der Verwendung des Standardschätzfehlers auf Basis der Regressionshypothese muss zusätzlich zu den Voraussetzungen, die für den Standardmessfehler genannt wurden, eine bivariate Normalverteilung der Messfehler und der wahren Werte vorliegen (Huber, 1973, S. 118), die jedoch unkorreliert sein müssen.

Prüfung der Voraussetzungen Die Voraussetzungen für Standardmess- und Standardschätzfehler sind schwer zu prüfen. In der Regel reicht es aus, wenn eine Normalverteilung des Testkennwerts vorliegt (Stelzl, 1993, S. 55). Man könnte also zumindest eine Inspektion des Histogramms des Testwerts vornehmen. Es ist wichtig zu erwähnen, dass, selbst wenn die Voraussetzungen verletzt sind oder nicht geprüft werden können, eine Absicherung gegenüber dem Messfehler erfolgen muss. Wichtiger als die strenge Erfüllung dieser Voraussetzungen ist es, dass Mittelwerte, Standardabweichungen und Reliabilitäten möglichst präzise geschätzt werden. Das heißt konkret, dass große und möglichst repräsentative Normstichproben sowie ausreichend lange Tests hoher psychometrischer Qualität vorliegen. Auch wenn man die den Vertrauensintervallen zugrunde liegenden Annahmen, wie vor allem die Unabhängigkeit der wahren Werte und der Messfehler sowie die Annahme konstanter Messfehler für jede Person, kritisieren kann (vgl. Krauth, 1995, S. 208 ff.), so führt doch kein Weg an der Absicherung der individuellen Testwerte mit einem Vertrauensintervall vorbei: **Vertrauensintervalle müssen in der Praxis angegeben werden – dies gehört zu den Qualitätsstandards für Gutachten, die unbedingt einzuhalten sind. Werden keine Vertrauensintervalle berechnet, ist ein Befund oder Gutachten nicht sachgemäß erstellt.** Bei aller Kritik wird durch das Vertrauensintervall dem Umstand Rechnung getragen, dass die Werte, die wir durch Tests erhalten, nicht perfekt gemessen werden und unter wiederholten Bedingungen nicht identisch ausfallen. Dabei ist die Anwendung des Standardmessfehlers für die Praxis meist völlig ausreichend.

4

Beispiel 4.6

Vertrauensintervalle um den beobachteten Wert

Beim 16-jährigen Markus soll abgeklärt werden, ob er in der Schule kognitiv über- oder unterfordert ist, da er in letzter Zeit schlechte Schulleistungen erbringt. Der 16-jährige Gymnasiast Markus erzielte im Test I-S-T 2000 R (Liepmann, Beauducel, Brocke & Amthauer, 2007, 2. Auflage) einen IQ von 111 im schlussfolgernden Denken. Der Mittelwert der IQ-Norm beträgt 100 und die Standardabweichung $\sigma_x = 15$. Cronbach-α wird für den Kennwert schlussfolgerndes Denken mit $\hat{\rho}_{tt} = 0.94$ angegeben. Die Sicherheitswahrscheinlichkeit soll zur Veranschaulichung einmal 95 Prozent ($z_{1\text{-}\alpha/2} = 1.96$) und einmal 90 Prozent ($z_{1\text{-}\alpha/2} = 1.64$) bei zweiseitiger Testung betragen. Der Testanwender möchte nun wissen, mit welcher Wahrscheinlichkeit das Vertrauensintervall den wahren Wert überdeckt.

Für 95 Prozent Sicherheitswahrscheinlichkeit:

Äquivalenzhypothese:

$$\sigma_{E_x} = 15 \cdot \sqrt{1-.94} = 3.67$$

$$VI_{u,o} = 111 \pm 3.67 \cdot 1.96 = 111 \pm 15 \cdot \sqrt{(1-.94)} \cdot 1.96$$

$$VI_{u,o} = 111 \pm 3.67 \cdot 1.96 = 103.8 - 118.2$$

Praxistipp

Beispieltext im Befund: Markus erzielte im I-S-T 2000 R (Liepmann, Beauducel, Brocke & Amthauer, 2007, 2. Auflage) im Vergleich zu 15- bis 25- jährigen Gymnasiasten einen durchschnittlichen bis überdurchschnittlichen Wert im schlussfolgernden Denken.

Beispielangabe im Anhang (des Befunds oder Gutachtens): Der I-S-T 2000 R (Liepmann, Beauducel, Brocke & Amthauer, 2007) wurde in der zweiten Auflage verwendet. Die Berechnung des Vertrauensintervalls erfolgte mit dem Standardmessfehler. Dabei wurde eine zweiseitige Testung durchgeführt. Als Sicherheitswahrscheinlichkeit wurde ein Wert von 95 Prozent gewählt sowie Cronbach-α als Reliabilitätsschätzung verwendet. Als Norm wurde die IQ-Norm 15- bis 25-jähriger Gymnasiasten gewählt, und dies ergab unter den gewählten Einstellungen folgendes Vertrauensintervall: 103.8 bis 118.2 bei einem beobachteten Wert von 111.

Regressionshypothese:

$$\sigma_{E_\tau} = 15 \cdot \sqrt{.94 \cdot (1-.94)} = 3.56$$

$$\hat{\tau}_v = .94 \cdot 111 + 100 \cdot (1-.94) = 110.3$$

$$VI_{u,o} = 110.3 \pm 3.56 \cdot 1.96 = 110.3 \pm 15 \cdot \sqrt{.94 \cdot (1-.94)} \cdot 1.96 = 103.3 - 117.3$$

Praxistipp

Beispieltext im Befund: Markus erzielte im I-S-T 2000 R (Liepmann, Beauducel, Brocke & Amthauer, 2007, 2. Auflage) im Vergleich zu 15- bis 25- jährigen Gymnasiasten einen durchschnittlichen bis überdurchschnittlichen Wert im schlussfolgernden Denken.

Beispielangabe im Anhang (des Befunds oder Gutachtens): Der I-S-T 2000 R (Liepmann, Beauducel, Brocke & Amthauer, 2007) wurde in der zweiten Auflage verwendet. Die Berechnung des Vertrauensintervalls erfolgte mit dem Standardschätzfehler nach vorheriger Schätzung des wahren Werts auf Basis der Regressionshypothese. Dabei wurde eine zweiseitige Testung durchgeführt. Als Sicherheitswahrscheinlichkeit wurde ein Wert von 95 Prozent gewählt sowie Cronbach-α als Reliabilitätsschätzung verwendet. Als Norm wurde die IQ-Norm 15- bis 25-jähriger Gymnasiasten gewählt, und dies ergab unter den gewählten Einstellungen folgendes Vertrauensintervall: 103.3 bis 117.3 bei einem beobachteten Wert von 111 und einem geschätzten wahren Wert von 110.3.

Für 90 Prozent Sicherheitswahrscheinlichkeit:

Äquivalenzhypothese:

$$\sigma_{E_x} = 15 \cdot \sqrt{1-.94} = 3.67$$

$$VI_{u,o} = 111 \pm 3.67 \cdot 1.64 = 111 \pm 15 \cdot \sqrt{(1-.94)} \cdot 1.64$$

$$VI_{u,o} = 111 \pm 3.67 \cdot 1.64 = 105.0 - 117.0$$

Praxistipp

Beispieltext im Befund: Markus erzielte im I-S-T 2000 R (Liepmann, Beauducel, Brocke & Amthauer, 2007, 2. Auflage) im Vergleich zu 15- bis 25- jährigen Gymnasiasten einen durchschnittlichen bis überdurchschnittlichen Wert im schlussfolgernden Denken.

Beispielangabe im Anhang (des Befunds oder Gutachtens): Der I-S-T 2000 R (Liepmann, Beauducel, Brocke & Amthauer, 2007) wurde in der zweiten Auflage verwendet. Die Berechnung des Vertrauensintervalls erfolgte mit dem Standardmessfehler. Dabei wurde eine zweiseitige Testung durchgeführt. Als Sicherheitswahrscheinlichkeit wurde ein Wert von 90 Prozent gewählt sowie Cronbach-α als Reliabilitätsschätzung verwendet. Als Norm wurde die IQ-Norm 15- bis 25-jähriger Gymnasiasten gewählt, und dies ergab unter den gewählten Einstellungen folgendes Vertrauensintervall: 105 bis 117 bei einem beobachteten Wert von 111.

Regressionshypothese:

$$\sigma_{E_\tau} = 15 \cdot \sqrt{.94 \cdot (1-.94)} = 3.56$$

$$\hat{\tau}_v = .94 \cdot 111 + 100 \cdot (1-.94) = 110.3$$

$$VI_{u,o} = 110.3 \pm 3.56 \cdot 1.64 = 110.3 \pm 15 \cdot \sqrt{.94 \cdot (1-.94)} \cdot 1.64 = 104.5 - 116.1$$

Praxistipp

Zusammenfassung. Der Proband mit einem IQ von 111 ist in allen dargestellten Fällen als durchschnittlich bis überdurchschnittlich intelligent einzustufen. Bei einer niedrigeren Reliabilität, beispielsweise von $\hat{\rho}_{tt} = 0.80$, können die Unterschiede in der Klassifikation auf Basis der Äquivalenz- und Regressionshypothese deutlicher ausfallen. Die Regressionshypothese ist dann aufgrund des schmaleren Vertrauensintervalls konservativer.

Würde man eine aufgrund der hohen Reliabilität in diesem Fall völlig ausreichende Sicherheitswahrscheinlichkeit von 80 Prozent wählen, würde auf Basis der Äquivalenzhypothese das Vertrauensintervall nach wie vor eine Klassifikation von „durchschnittlich bis überdurchschnittlich“ ergeben, während auf Basis der Regressionshypothese das Vertrauensintervall eine Klassifikation von „durchschnittlich“ ergeben würde. Es gibt an dieser Stelle viele Möglichkeiten, die Breite des Vertrauensintervalls zu beeinflussen. Wichtig ist, dass die Wahl der Einstellungen gut begründet werden kann.

4.8.2 Bedeutsamkeit von Untertestdifferenzen

Wie groß müssen die Leistungsunterschiede eines Probanden in zwei Tests sein, um als abgesichert gelten zu können?

Zur Beantwortung der Frage, ob sich die Leistungen eines Probanden in zwei Untertests oder Tests unterscheiden, wird die **interne Konsistenz** als **Reliabilitätsschätzer** zur einzelfalldiagnostischen Auswertung benötigt, da lediglich eine Aussage über den momentanen Status und daher keine Prognose verlangt wird.

Voraussetzungen Im Folgenden ist die Absicherung von Testwertdifferenzen unterschiedlicher Skalen bei einer Person dargestellt. Voraussetzung bei unterschiedlichen Tests ist, dass gleiche Normwerte vorliegen, beispielsweise verwenden beide Tests eine IQ-Norm. Diese wird im nächsten Abschnitt beschrieben. Sind die Reliabilitäten relativ ähnlich, wird die kritische Differenz nach der nachfolgend dargestellten Formel bestimmt.

Die hier dargestellten Formeln sind dann anzuwenden, wenn die Reliabilitäten der beiden Messwerte ähnlich hoch ausfallen. Weichen beide Reliabilitäten stark voneinander ab, sollte eine Tau-Normierung durchgeführt werden (siehe Kapitel 4.8.3) und eine messfehlerkritische und/oder schätzfehlerkritische Absicherung vorgenommen werden.

Ursachen für Testwertdifferenzen Neben der Frage, ob eine beobachtete Messwertdifferenz auch deshalb aufgetreten sein könnte, weil beide Messwerte messfehlerbehaftet sind, kann auch die Frage gestellt werden, ob eine beobachtete Differenz unter Berücksichtigung der Korrelation (Validität) beider Skalenwerte selten in der Normstichprobe auftritt. Huber (1973) bezeichnet die letztere Frage auch als Frage der **diagnostischen Valenz** oder **valenzkritische Absicherung**. Voraussetzung dafür ist, dass beide Testwerte bivariat normalverteilt sind (siehe Bühner & Ziegler, 2009, S. 606).

Absicherung als zweistufiger Prozess Huber (1973) sieht die Frage nach den Ursachen für das Auftreten von Messwertdifferenzen generell als zweistufigen Prozess an. Zunächst wird gefragt, ob der beobachtete Unterschied auf **Messfehler** zurückgeführt werden kann. Kann der beobachtete Unterschied nicht alleine auf Messfehler zurückgeführt werden, schließt sich die Frage an, ob die beobachtete Messwertdifferenz **in der Normstichprobe besonders selten vorkommt** und damit bereits klinisch auffällig ist. Um diese Fragen zu klären, werden die folgenden Schritte durchgeführt:

- Bildung der kritischen Differenz
- Vergleich der kritischen Differenz mit der beobachteten Differenz

Anwendung folgender Entscheidungsregeln:

- $D_{kritisch} \geq D_{beobachtet}$ → Unterschied kann auf Messfehler der Messwerte zurückgeführt werden (messfehlerkritisch) oder der gefundene Unterschied ist diagnostisch nicht bedeutsam (valenzkritisch).
- $D_{kritisch} < D_{beobachtet}$ → Unterschied kann **nicht** alleine auf Messfehler der Messwerte zurückgeführt werden (messfehlerkritisch) oder der gefundene Unterschied **ist** diagnostisch bedeutsam (valenzkritisch).

Messfehlerkritische Unterschiede (Standardmessfehler)

Sind die Reliabilitäten der Tests nur leicht verschieden:

$$D_{\text{krit.intra}} = z_{1-\frac{\alpha}{2}} \cdot \sigma_x \cdot \sqrt{2-(\rho_{tt1}+\rho_{tt2})}$$

Sind die Reliabilitäten der Tests gleich (z.B. bei Retestung mit dem gleichen Test), kann man folgende vereinfachte Formel verwenden:

$$D_{\text{krit.intra}} = z_{1-\frac{\alpha}{2}} \cdot \sigma_x \cdot \sqrt{2\cdot(1-\rho_{tt})}$$

- $D_{krit.intra}$ = kritische Differenz intraindividuell (Unterschiede in den Leistungen einer Person)
- σ_x = Standardabweichung des Tests
- ρ_{tt_1}= Reliabilität des Messwerts von Untertest 1
- ρ_{tt_2}= Reliabilität des Messwerts von Untertest 2
- $z_{1-\alpha/2}$= kritischer z-Wert der Standardnormalverteilung für eine bestimmte Irrtumswahrscheinlichkeit α bei zweiseitiger Testung

Exkurs 4.8

Vorsicht! Nicht alles, was man testen kann, sollte man testen …

Wichtig ist in diesem Zusammenhang, dass diese Formeln nicht angewendet werden können, wenn ein Untertest- und ein Gesamttestwert oder zwei Skalen- oder Untertestkennwerte, die eine gewisse Anzahl an Untertests oder Items teilen, auf Unterschiedlichkeit geprüft werden sollen (Stelzl, 1982). **In einem solchen Fall ist die Annahme unkorrelierter Messfehler verletzt.** Nehmen wir als Beispiel den Messwert zum schlussfolgernden Denken aus dem I-S-T 2000 R und die verbale Intelligenz aus dem I-S-T 2000 R: In die Berechnung des Messwerts zum schlussfolgernden Denken geht der Messwert der verbalen Intelligenztestskala mit ein. Damit sind die Messfehler des Kennwerts der verbalen Intelligenztestskala im Messwert zum schlussfolgernden Denken enthalten. Die Annahme unkorrelierter Messfehler ist damit verletzt.

Unterschiede in der diagnostischen Valenz (Standardmessfehler)

$$D_{\text{krit.intra}} = z_{1-\frac{\alpha}{2}} \cdot \sigma_x \cdot \sqrt{1-\rho_{12}^2}$$

- $D_{krit.intra}$= kritische Differenz intraindividuell (Unterschiede in verschiedenen Leistungen einer Person)
- $z_{1-\alpha/2}$= kritischer z-Wert der Standardnormalverteilung für eine bestimmte Irrtumswahrscheinlichkeit α bei zweiseitiger Testung
- σ_x = Standardabweichung des Tests (unter der Bedingung $\sigma_1 = \sigma_2$)
- ρ^2_{12}= Determinationskoeffizient (quadrierte Korrelation) zwischen Test 1 und Test 2

Es muss gelten: Standardabweichungen beider Tests sind gleich $(\sigma_1 = \sigma_2)$, was bei der Verwendung von denselben Normwerten gegeben ist. Weiterhin muss eine bivariate Normalverteilung der Messwerte vorliegen und die Reliabilität der Messwerte gleich sein.