4. Die Klassische Testtheorie (KTT)

Stellen wir uns vor, die Messung sei nun durchgeführt, das Messinstrument wäre bei einer Reihe von Personen angewandt worden. Das Resultat der Messung ist nun eine Menge von Zahlen. Eine offene Frage bleibt allerdings wie gut die Messung nun eigentlich gelungen ist. Bei der wissenschaftlichen Anwendung von Tests müssen mehrere wissenschaftliche Gütekriterien erfüllt sein, dass man von einer guten Messung sprechen kann.

Die Testgütekriterien sind:

1. Objektivität: Unabhängigkeit der Testergebnisse vom Untersucher. Man unterscheidet die Objektivität der Testdurchführung, der Auswertung und der Interpretation. Ohne Objektivität gibt es keine Zuverlässigkeit (Reliabilität).
   Probleme entstehen z.B. durch Versuchsleitereffekte nach Rosenthal (1976).

2. Reliabilität: Zuverlässigkeit des Tests, bzw. Genauigkeit, mit der gemessen wurde, oder Stabilität der Messung über die Zeit. Man unterscheidet Retest-Reliabilität, Paralleltest-Reliabilität und

interne Konsistenz (quasi parallele Tests). Die Reliabilität ist die Voraussetzung, aber nicht die hinreichende Bedingung für die Validität des Tests.

3. Validität: Gültigkeit des Tests. Man muss die Frage stellen: "Misst der Test das, was er vorgibt zu messen?" Grad der Genauigkeit, mit der der Test eine interessierende latente Eigenschaft (z.B.: Persönlichkeit) bzw. eine beliebige Kriteriumsvariable zu messen gestattet. Man kann zwischen inhaltlicher Validität, Kriteriumsvalidität und Konstrukt-Validität unterscheiden. Die Validität eines Tests ist nicht nur von der Reliabilität dieses Tests abhängig, sondern auch von der Reliabilität der Messung der Kriteriumsvariable.

Nebenkriterien der Güte von Tests sind (pragmatischer Aspekt) nach Lienert:

1. Normiertheit: Sind Aussagen über die "Normalbevölkerung" möglich?

2. Vergleichbarkeit: Können die Ergebnisse mit anderen Tests verglichen werden? Welche Stichprobe bildet die Vergleichsgruppe? Sind die Ergebnisse bzgl. Alter, Geschlecht und anderer Variablen vergleichbar?

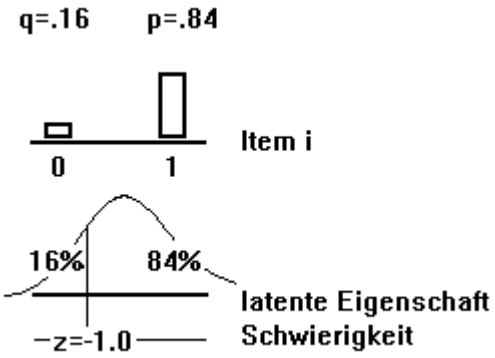
3. Oekonomie: Wie teuer und zeitaufwendig ist die Testanwendung?

4. Nützlichkeit: Bringt der Tests einen Nutzen für den Testanwender bzw. für den Getesteten?

Diese Anforderungen an einen Test bzgl. Objektivität, Reliabilität und Validität gelten ebenso für die einzelnen Aufgaben (Items) aus denen sich der Test zusammensetzt. Zusätzlich gibt es aber noch drei weitere Aufgabengütekriterien:

1. Itemschwierigkeit: Anzahl richtiger Lösungen durch die Gesamtzahl der Antworten auf einem Item (eigentlich Itemleichtigkeit).

Abbildung 4: Itemschwierigkeit bei dichotomen und kontinuierlichen Items



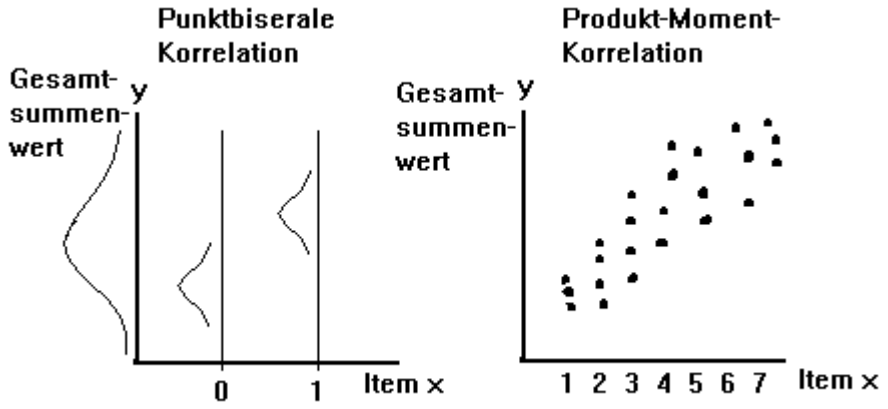
Legende: Der obere Teil der Abbildung zeigt ein dichotomes Item. Die Antworten sind kodiert mit 0: "Aufgabe nicht gelöst" und 1: "Aufgabe gelöst". Es werden die Anzahl der richtigen Lösungen durch die Gesamtzahl dividiert. Daraus ergibt sich hier eine Itemschwierigkeit von p=.84.

Der untere Teil zeigt ein kontinuierliches Items. Die Personen die z-Werte von grösser als -1.0 erreichen haben das Item positiv angekreuzt, somit in unserem Sinne gelöst.

Abbildung 4 zeigt eine Itemschwierigkeit von p=.84. Das heisst, dass 84 von 100 Personen das Item lösen konnten. Bei kontinuierlichen Items wird eine Normalverteilung der Antworthäufigkeiten angenommen. Die Itemschwierigkeit p=.84 ist die Fläche unter der Verteilung die 84% der Gesamtfläche ausmacht.
Hohe Schwierigkeit bedeutet das Item ist leicht zu lösen. Gesucht werden Items mit mittlerer Schwierigkeit um p=.50 (Voraussetzung für hohe Trennschärfe).

2. Trennschärfe: Wie gut trennt die einzelne Aufgabe die Probanden mit einem hohen Testscore von denen mit einem niedrigen Testscore. Berechnet wird die Trennschärfe durch die Korrelation zwischen Testscore und Aufgabenbeantwortung, also die Korrelation des Einzelitems mit der Skala. Personen mit hohen Summenwerten im Test sollten eine Aufgabe eher lösen können, als Personen mit niedrigen Werten.

Abbildung 5: Zusammenhang zwischen Gesamtsummenwert und Beantwortung von dichotomen und kontinuierlichen Items



Legende: Der linke Teil der Abbildung stellt eine positive Punktbiserale Korrelation zwischen Itembeantwortung und Gesamtsummenwert dar. Die Antworten sind kodiert mit 0: "Aufgabe nicht gelöst" und 1: "Aufgabe gelöst". Für eine Person im unteren Bereich des Gesamtsummenwertes ist es wahrscheinlicher, dass sie das Item nicht lösen wird.
Der rechte Teil der Abbildung zeigt eine positive Produkt-Moment-Korrelation zwischen kontinuierlichem Item und Gesamtsummenwert. Je höher eine Person auf dem Item ankreuzt, umso höher ist auch ihr Gesamtsummenwert.

Abbildung 5 zeigt, dass bei dichotomen Items die Häufigkeiten der falschen bzw. der richtigen Antworten auf einem Item mit der Häufigkeitsverteilung der Antworten auf allen Items (Gesamtsummenwert) zusammenhängen müssen.
Gewünscht werden hohe Trennschärfen $r_{ig\text{-}i} \geq .30$.

3. Homogenität: Wenn die Aufgaben dieselbe Eigenschaft messen, sind sie homogen. Nur homogene Items sollten zu einer Skala (Gesamtsummenwert) zusammengefasst werden. Die Trennschärfe ist der Indikator für die Homogenität (s. u. bei der Faktorenanalyse).

4.1. Die Annahmen der klassischen Testtheorie

Die klassische Testtheorie geht von dem Denkmodell aus, jede Messung sei beliebig oft zu wiederholen (Problem: Phänomene in der Psychologie ändern sich allein schon durch die Messung z.B.: Reaktivität der Messung; Habituation. Zudem können viele Phänomene in der Psychologie nur einmal gemessen werden so zum Beispiel der erste Eindruck).

X_{vil} . . . beobachteter Messwert der Person v im Test i bei der l-ten Vorgabe
T_{vi} . . . wahrer Wert der Person v im Test i (griechisch: Tau)
f_{vil} . . . Fehler bei der l-ten Messung

$$X_{vil} = T_{vi} + f_{vil}$$

Annahme: Der Testwert setzt sich aus wahrem Wert und Fehler zusammen.

Bei beliebig ofter Wiederholung der Messung ist F_{vi} eine Zufallsvariable mit den Ausprägungen $\{f_{vi1}, f_{vi2}, \ldots, f_{vil}\}$, die zufällig mal nach oben, mal nach unten schwanken kann. Der wahre Wert T_{vi} sei konstant. Damit unterliegt auch der Testwert X_{vi} bei mehrmaliger Wiederholung der Messung den Zufallsschwankungen.

$$X_{vi} = T_{vi} + F_{vi}$$

Exkurs: Definition eines Erwartungswertes: Der Erwartungswert ist der "erstrebte" Wert der Realisation einer Zufallsvariable aus einer Stichprobe. Es ist der Wert der am wahrscheinlichsten auftritt. Bsp.: Bei einer Normalverteilung einer Zufallsvariable ist der Erwartungswert einer Ziehung der wahrscheinlichste Wert, nämlich der Mittelwert.

$E(F_{vi}) = 0$ Annahme: Der Erwartungswert des Fehlers bei mehrmaliger Messung ist Null.

$E(X_{vi}) = T_{vi}$	Merke: Der wahre Wert ist der Erwartungswert des Testwertes.

$$\sigma^2(X_{vi}) = \sigma^2(F_{vi})$$

Varianzen sind quadrierte Standardabweichungen, die mit dem griechischen Buchstaben Sigma (σ) bezeichnet
Die Variabilität der Gesamtvarianz kommt durch die Fehlervarianz zustande, da T_{Vi} als konstant angenommen wurde.

Führen wir wiederholte Messungen an zufällig aus einer Population ausgewählten Personen durch, ist auch T eine Zufallsvariable mit den Ausprägungen $\{T_1, T_2, \ldots, T_V\}$ (Index i lassen wir weg, da es sich um einen Test handeln soll)

$X = T + F$

4.2. Die Axiome der klassischen Testtheorie

1. $E(F) = 0$

2. $\rho(F,T) = 0$

3. $\rho(F_1,T_2) = 0$

4. $\rho(F_1,F_2) = 0$

Der griechische Buchstabe Rho (ρ) bezeichnet die Korrelation zweier Variablen in der Population.
Axiom 2 drückt aus, dass Fehler und wahrer Wert nicht zusammenhängen. Entsprechend darf auch eine Fehlerkomponente 1 (aus der Messung 1) nicht mit dem wahren Wert 2 (aus Messung 2) zusammenhängen (Axiom 3). Ebenso sollen die Fehlerkomponenten aus zwei Messungen nicht miteinander korrelieren.
Die Axiome drücken letztlich aus, dass alle Messfehler rein zufällig sind.
Problem: Axiom 2 legt nahe, dass in allen Bereichen der Skala gleich gut gemessen werden kann. Die Frage bleibt allerdings, ob im Extrembereich derselbe Messfehler festzustellen ist, wie im mittleren Bereich.

Ableitungen aus den Axiomen:

$E(X) = E(T)$

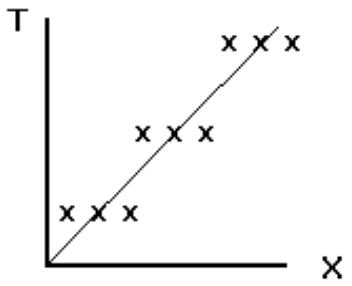
$\sigma^2(X) = \sigma^2(T) + \sigma^2(F) + 2 * \sigma^2(T) * \sigma^2(F) * r_{TF}$

Die Varianz einer Summe (X = T + F) ist die Varianz der einzelnen Summanden plus 2 mal die Kovarianz. Der arabische Buchstabe r steht für die Korrelation zweier Merkmale in einer Stichprobe. Nach dem Axiom 2 ist $r_{TF} = 0$, so dass die Kovarianz wegfällt, und es gilt:

$$\sigma^2(X) = \sigma^2(T) + \sigma^2(F)$$

Merke: Die Varianz der Testwerte ist gleich der Summe aus der Varianz der wahren Werte und der Messfehlervarianz.

Abbildung 6: Regression der Messwerte X auf die wahren Werte T



Legende: Auf der x-Achse sind die Messwerte X aufgetragen, auf der y-Achse die wahren Werte T. Die Korrelation der Messwerte und der wahren Werte ist 1. Die Schwankungen um die Regressionsgerade kommen durch den Fehler zustande.

Merke: Da der Erwartungswert von X gleich dem wahren Wert ist, ist die Regression von X auf T linear mit dem Anstieg 1. Die Varianz der Testwerte um die Regressionsgerade ist konstant.

4.3. Reliabilität

4.3.1. Definition von Reliabilität

Reliabilität ist der Anteil der wahren Varianz an der Testvarianz (Gesamtvarianz). Sie drückt aus, wie gross der Fehleranteil bei der Messung war, man spricht deshalb auch von der Zuverlässigkeit der Messung. Die Reliabilität entspricht dem Quadrat der Korrelation des Rohwertes mit dem wahren Wert:

$$r_{tt} = \frac{\sigma^2(T)}{\sigma^2(X)} = \rho^2(X,T)$$

Vorsicht: Die Reliabilität ist allgemein mit r_{tt} bezeichnet, sie ist aber keine reine Korrelation, sondern das Quadrat einer Korrelation, wie z.B. der Determinationskoeffizient r^2 (gemeinsame Varianz zweier Variablen).

(Andere Varianzverhältnisse sind in der Regression: $R^2 = \frac{SS_{reg}}{SS_{tot}}$;

oder in der Varianzanalyse: $Eta^2 = \frac{SS_{zw}}{SS_{tot}}$).

Da Varianzen stets positiv sind (da quadriert), und die wahre Varianz stets kleiner ist als die Testvarianz ergibt sich:

$$0 \leq r_{tt} \leq 1$$

Da der wahre Wert nicht beobachtbar ist, kann die Reliabilität weder aufgrund des Varianzverhältnisses noch aufgrund der Korrelation bestimmt werden. Die praktische Bestimmung der Reliabilität erfolgt über einen Trick: der Paralleltest.

4.3.2. Paralleltests

Definition: Paralleltests messen dieselbe latente Dimension und haben denselben Messfehler:

$X = T + F$ $\qquad$ $X' = T + F'$

Zwei unabhängige Messungen heissen parallel, wenn gilt:

1. $E(X) = E(X') = E(T)$

2. $\sigma^2(X) = \sigma^2(X')$

Paralleltests haben gleiche Erwartungswerte und gleiche Varianz.

Denkmodell: Könnte man zweimal dasselbe Merkmal völlig fehlerfrei messen, wäre $r_{tt} = 1$. Man würde zwei wahre Werte miteinander korrelieren. Schätzen liesse sich die Reliabilität, wenn man in einem Test zwei wahre Werte bestimmen könnte, und diese miteinander korreliert (durch Paralleltests).

Mathematisch muss man nun zeigen, dass die Korrelation der parallelen Messungen X_1 und X_2 sich tatsächlich in die Definitionsformel der Reliabilität überführen lässt. Ausgehen muss man von der Formel der Pearson-Produkt-Moment-Korrelation:

$$r_{X_1X_2} = \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1) * \sum_{i=1}^{n}(X_{2i} - \bar{X}_2)}{s_{X_1} * s_{X_2} * (n-1)}$$

Zur Vereinfachung lassen wir den Index i weg. Zudem kann man die Differenzen vom Mittelwert noch einfacher schreiben:

$x_1 = X_1 - \bar{X}_1$; $t_1 = T_1 - \bar{T}_1$; $f_1 = F_1$

$x_2 = X_2 - \bar{X}_2$; $t_2 = T_2 - \bar{T}_2$; $f_2 = F_2$

$$r_{X_1X_2} = \frac{\sum_{i=1}^{n} x_1 * \sum_{i=1}^{n} x_2}{s_{x_1} * s_{x_2} * (n-1)}$$

Wir können einsetzten: $x_1 = t_1 + f_1$ und $x_2 = t_2 + f_2$

$$r_{X_1X_2} = \frac{\sum_{i=1}^{n} (t_1 + f_1) * \sum_{i=1}^{n} (t_2 + f_2)}{s_{x_1} * s_{x_2} * (n-1)}$$

$$r_{X_1X_2} = \frac{\sum_{i=1}^{n} (t_1\, t_2 + t_1\, f_2 + t_2\, f_1 + f_1\, f_2)}{s_{x_1} * s_{x_2} * (n-1)}$$

$$r_{X_1X_2} = \frac{\sum_{i=1}^{n} t_1\, t_2 + \sum_{i=1}^{n} t_1\, f_2 + \sum_{i=1}^{n} t_2\, f_1 + \sum_{i=1}^{n} f_1\, f_2}{s_{x_1} * s_{x_2} * (n-1)}$$

$$r_{X_1X_2} = \frac{\sum_{i=1}^{n} t_1\, t_2}{s_{x_1} * s_{x_2} * (n-1)}$$

Wenn zwei äquivalente Messungen vorliegen, messen beide Tests dieselben wahren Werte. Zudem sollen noch die Standardabweichungen gleich sein, dann gilt:

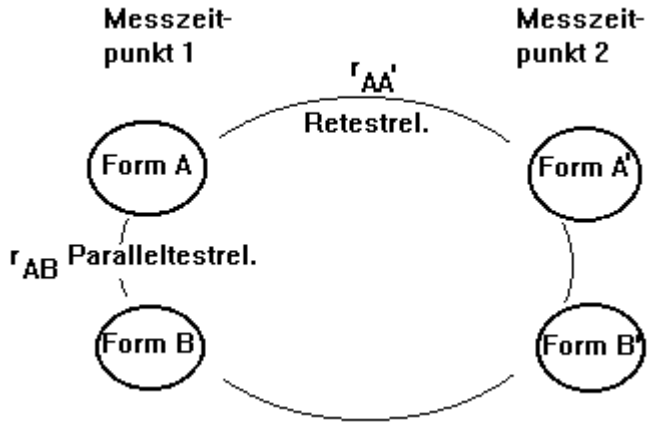
$$r_{X_1X_2} = \frac{\sum_{i=1}^{n} t^2}{s_x^2 * (n-1)} = \frac{s_t^2}{s_x^2} = \frac{s_T^2}{s_X^2} = r_{tt}$$

Merke: Die Reliabilität eines Tests ist gleich der Korrelation zu seinem Paralleltest.

Vorsicht: Hier ist die einfache Korrelation gemeint, nicht wie oben das Quadrat der Korrelation.

4.3.3. Praktische Methoden zur Bestimmung der Reliabilität

Abbildung 7: Praktische Reliabilitätsbestimmung



Legende: Zur parallelen Messung eines Merkmals können zu einem Messzeitpunkt zwei Formen eines Tests vorgelegt werden (Form A und Form B). Die Korrelation beider Messungen entspricht der Paralleltest-Reliabilität. Wenn ein Test zu zwei Messzeitpunkten vorgelegt wird (Form A und Form A') spricht man von der Retest-Reliabilität.

1. Retest-Reliabilität: Die Messung mit dem Test wird zweimal an derselben Stichprobe durchgeführt. Man erhält so zwei Schätzungen für die wahren Werte jeder Person (auch Stabilitätskoeffizient genannt, da die Messungen über zwei Zeitpunkte stabil bleiben sollen).

2. Paralleltest-Reliabilität: Man konstruiert zum vorhandenen Test eine vollkommen äquivalente Parallelform. Auch hier erhält man zwei Schätzungen der wahren Werte.

3. Odd-Even- oder Split-Half-Reliabilität: Man halbiert den vorhandenen Test in zwei Testhälften (gerade-ungerade oder Anfang vs. Ende). (Testhalbierungsmethode)

4. Interne Konsistenz: Man kann den Test in beliebig viele Teile aufteilen, bis auf die Ebene der einzelnen Items. Durch eine verallgemeinerte Formel wird dann die Reliabilität anhand der n Testteile geschätzt. Besonderes bei wenigen Items wird diese Methode oft angewandt.

4.3.4. Formeln zur praktischen Bestimmung der Reliabilität

Häufig wird die Reliabilität über die Testhalbierungsmethode (Punkt 3) bestimmt. Man könnte nun einfach die beiden Testhälften miteinander korrelieren um die Reliabilität zu schätzen. Durch die Halbierung haben wir allerdings einen gewissen Informationsverlust, was zu einer Unterschätzung der Reliabilität führt. Eine Korrekturmöglichkeit bietet die Spearman-Brown Formel:

$$r_{ttn} = \frac{n * r_{tt}}{1 + (n - 1) * r_{tt}} \quad \text{bei 2 Testhälften:} \quad r_{ttn} = \frac{2 * r_{tt}}{1 + r_{tt}}$$

r_{ttn} ... Reliabilitätsschätzung nach "Aufwertung"
r_{tt} ... Korrelation der Testhälften
n ... Anzahl der Testverlängerungen

Diese Formel kann auch verwendet werden, um herauszufinden, wie oft man einen Test verlängern muss (gesucht wird n), um eine bestimmte Reliabilität zu erzielen. Bsp.: Ich habe einen Test mit r_{tt} = .75, möchte aber ein r_{ttn} = .90. Wie gross ist n?

$$n = \frac{r_{ttn} * (1 - r_{tt})}{r_{tt} * (1 - rtt_n)} = \frac{.90 * (1 - .75)}{.75 * (1 - .90)} = 3$$

Ich muss den Test 3 mal verlängern.

Zur Schätzung der internen Konsistenz (4.) wird auch sehr häufig die Formel von Cronbach (Cronbach Alpha) verwendet, die nicht von der Korrelation der Testhälften ausgeht, sondern von den Varianzen der Items und der Testvarianz:

$$\alpha = \frac{m}{m - 1} * \left(1 - \frac{\sum_{j=1}^{m} s_j^2}{s_x^2} \right)$$

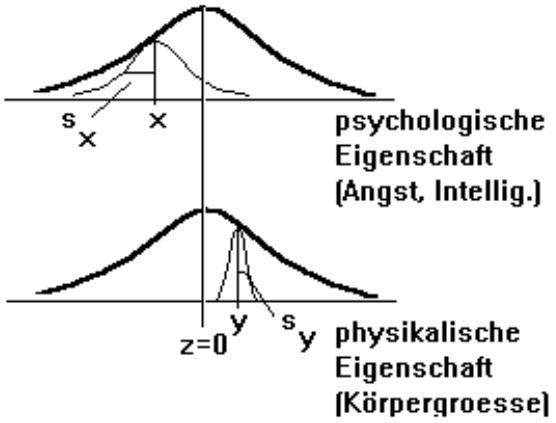
s_j^2 ... Varianz des Items j (j = 1, 2, ... , m)
m ... Anzahl der Items des Tests
s_x^2 ... Varianz des Tests

Diese beiden Formeln werden hier angegeben, da sie in Computerprogrammen oft Verwendung finden. Sie sind mathematisch ineinander überführbar, und zeigen nur bei extremen Daten Unterschiede. Die Spearman-Brown Formel wird in der Literatur eher zitiert. Sie schätzt aber nur bei wirklich parallelen Testteilen korrekt. Deshalb ist die konservativere Schätzung des Cronbach Alpha generell vorzuziehen.

Ein Spezialfall des Cronbach Alpha ist die "Kuder-Richardson-20" Formel (auch $Alpha_{20}$) für dichotome Items (0/1 - Antworten), die hier nicht weiter dargestellt werden soll.

4.3.5. Fehler bei der Messung: der Standardmessfehler

Abbildung 8: Variabilität bei der Messung psychologischer und physikalischer Eigenschaften



Legende: Der obere Teil der Abbildung zeigt die Variation einer psychologischen Eigenschaft in der Population (fette Verteilung). Die mehrmalige Messung einer Person x ergibt eine Verteilung mit der Sandardabweichung s_x (dünne Verteilung). Der untere Teil zeigt eine physikalische Eigenschaft mit entsprechender Variabilität in der Population wie die psychologische Eigenschaft (fette Verteilung). Die mehrmalige Messung der Person y führt zu einer kleineren Standardabweichung s_y (dünne Verteilung).

Abbildung 8 zeigt, dass wir unterschiedlich grosse Variabilitäten bei der Messung erhalten können. Es werden somit unterschiedlich grosse Fehler bei der Messung gemacht, obwohl in diesem Beispiel dieselbe Variabilität der Eigenschaften vorliegt.

Definition: Standardmessfehler:
Der Standardmessfehler ist die Standardabweichung durch die Wiederholung der Messung.

Der Fehler sollte möglichst klein sein in Bezug zur ganzen Variabilität. Wenn der Standardfehler über die ganze Variabilität geht, ist die Messung vollkommen unreliabel. Der Standardmessfehler hat grosse praktische Bedeutung.

4.4. Die Bedeutung der Reliabilität und des Standardmessfehlers in der diagnostischen Praxis

Praktische diagnostische Fragestellungen können sein:

1. In welchem Bereich liegt der "wahre Wert" einer getesteten Person? (Konfidenzintervall)
2. Unterscheiden sich die Leistungen einer Person in zwei Tests signifikant? (intraindividuelle Differenz)
3. Unterscheiden sich die Leistungen zweier Personen im gleichen Test? (interindividuelle Differenz)

zu 1.

$$r_{tt} = \frac{\sigma^2(T)}{\sigma^2(X)} = \frac{\sigma^2(X) - \sigma^2(F)}{\sigma^2(X)} = 1 - \frac{\sigma^2(F)}{\sigma^2(X)}$$

Dies ist die Formel der Reliabilität für Populationen. Für Stichproben können wir das Sigma (σ) durch ein s ersetzten und schreiben:

$$r_{tt} = 1 - \frac{s_F^2}{s_X^2}$$

Die Fehlervarianz ist die quadrierte Abweichung vom "wahren Wert", die durch die wiederholte Messung zustande kam. Durch Umformung der Formel lässt sie sich schätzten:

$$s_F^2 = s_X^2 \ (1 - r_{tt})$$

Will man keine Varianz, sondern eine Standardabweichung des Fehlers (man spricht vom Standardmessfehler), muss man aus dieser Formel die Wurzel ziehen:

$$s_F = s_X * \sqrt{1 - r_{tt}}$$

Merke: Der Standardmessfehler setzt sich zusammen aus der Standardabweichung des Tests mal der Wurzel aus der Unzuverlässigkeit.

Anwendungsbeispiel

Wir haben einen Intelligenztest durchgeführt. Die Standardabweichung des Tests ist $s_X = 10$; die Reliabilität beträgt $r_{tt} = .84$. Unsere Versuchsperson erreichte einen Testwert von X = 120. Der wahre Wert T der Person liegt bei einer statistischen Sicherheit von 95% (z-Wert = 1,96) in folgendem Bereich:

$$T = X \pm Z_{krit} * s_F$$

s_F kann nun bestimmt werden:

$$s_F = s_X * \sqrt{1 - r_{tt}} = 10 * \sqrt{1 - .84} = 4$$

Für die Bestimmung des Konfidenzintervalls ergibt sich dann:

$$T = 120 \pm 1{,}96 * 4 = 120 \pm 7{,}84$$

Mit einer statistischen Sicherheit von 95% liegt die wirkliche Leistung des Probanden in dem Bereich von 112,16 bis 127,84 IQ-Punkten.

zu 2.
Der Standardfehler intraindividueller Differenzen:

$$s_{F\,(Intra)} = s * \sqrt{2 - (r_{11} + r_{22})}$$

s ... Standardabweichung der beiden Tests
r_{11} ... Reliabilität des Tests 1
r_{22} ... Reliabilität des Tests 2

Man kann nun berechnen, wie gross die Differenz zweier Messwerte einer Person sein muss, um signifikant zu werden.

Anwendungsbeispiel:

Für zwei Tests liegen T-Werte vor (Mittelwert 50; Standardabweichung s=10). Test 1 hat eine Reliabilität von $r_{11} = .90$; für Test 2 ist $r_{22} = .74$.

$s_{F\ (Intra)} = s * \sqrt{2 - (r_{11} + r_{22})} = 10 * \sqrt{2 - (.90 + .74)} = 6$

Um wieviel Punkte muss die Leistung einer Person in Test 2 höher sein als in Test 1, damit die Differenz auf dem 5%-Niveau signifikant ist? Aus der z-Verteilung ermitteln wir einen kritischen z-Wert bei 5% und einseitiger Testung von 1,65. Damit ist

$X_2 - X_1 = Z_{krit} * s_{F(Intra)} = 1{,}65 * 6 = 10$

liegt der Wert einer Person in Test 2 um mehr als 10 Punkte höher als in Test 1 hatte die Person eine signifikant höhere Leistung im zweiten Test.

zu 3.
Der Standardfehler interindividueller Differenzen:

$$s_{F\ (Inter)} = s * \sqrt{2 * (1 - r_{tt})}$$

s ... Standardabweichung des Tests
r_{tt} ... Reliabilität des Tests

Wie gross muss der Unterschied zweier Probanden im gleichen Test sein, um auf einem bestimmten Niveau signifikant zu sein?

Anwendungsbeispiel:

Ein Test habe eine Standardabweichung von s = 20 und eine Reliabilität von r_{tt} = .92. Bei wieviel Punkten ist die Leistung zweier Probanden auf dem 10%-Niveau signifikant?

$s_{F\ (Inter)} = s * \sqrt{2 * (1 - r_{tt})} = s * \sqrt{2 * (1 - .92)} = 8$

Aus der z-Verteilung ermitteln wir einen kritischen z-Wert bei 10% und einseitiger Testung von 1,28. Damit ist

$X_A - X_B = Z_{krit} * s_{F(Inter)} = 1{,}28 * 8 = 10$

Wenn Person A 10 Testpunkte mehr erzielt, als Person B, ist der Leistungsunterschied mit 90 % Sicherheit überzufällig.

4.5. Die Reliabilität von Differenzwerten

Die bisherige Berechnung der Reliabilität bezog sich auf die Verrechnung von Absolutwerten bzw. Rohwerten. Wenn wir Veränderungen erfassen wollen, z. B. ob sich die Angstwerte eines Klienten nach der Therapie verändert haben, so können wir auch Differenzen miteinander vergleichen. Die Frage ist allerdings, ob die Veränderungen auch reliabel erfasst worden sind.
Um dies zu prüfen, können auch für Differenzen Reliabilitäten errechnet werden. Frage ist hier, wie zuverlässig war die Messung der Differenz. Die angegebene Formel (r_{ttdiff}) gilt für die Differenz zweier Tests bzw. Skalen oder für die Vorgabe eines Tests zu zwei Messzeitpunkten.

$$r_{tt\ diff} = \frac{r_{tt1} + r_{tt2} - 2 * r_{12}}{2 * (1 - r_{12})}$$

gültig für $r_{12} \neq 1$

$r_{tt\ diff}$ ... Reliabilität der Differenz
r_{tt1} ... Reliabilität von Test 1 (oder Zeitpunkt 1)
r_{tt2} ... Reliabilität von Test 2 (oder Zeitpunkt 2)
r_{12} ... Interkorrelation von Test 1 und Test 2
(oder zwischen den Zeitpunkten)

Diese Formel ist äusserst problematisch. Für $r_{12} = 1$ ist sie nicht definiert (man darf nicht durch Null teilen). In der Literatur wird häufig darauf hingewiesen, Differenzwerte seien viel unreliabler als Testscores. Dies geht u.a. auch auf diese Formel zurück. Je grösser r_{12} wird, umso kleiner wird die Reliabilität der Differenz (da [-2 * r_{12}] im Zähler steht). Es sollte klar sein, dass sehr stabile Merkmale wenig Varianz für Veränderungen zulassen. Nur wenn Vortest und Nachtest gering korrelieren besteht die Möglichkeit, dass die Differenzwerte, die diese Veränderung abbilden sollen genügend "wahre Varianz" besitzten. Die Reliabilität war definiert als Anteil der wahren Varianz zur Gesamtvarianz. Eine geringe Korrelation von Vortest und Nachtest widerspricht aber der Retestreliabilität, man kann damit nicht mehr davon ausgehen, dass dasselbe gemessen wurde.

In der neueren Literatur werden mehr und mehr "veränderungssensitive Messinstrumente" entwickelt. Zudem hat Wittmann (1985, S. 50 ff.) eine neue Messtheorie für die Reliabilität von Differenzen entwickelt, die hier nicht näher dargestellt werden soll. Allgemein wird empfohlen Rohwerte und Differenzwerte zu analysieren, wenn man Veränderungen abbilden will.

4.6. Validität

4.6.1. Definition von Validität

Validität bedeutet Gültigkeit. Sie bestimmt das Ausmass in dem der Test das misst, was er zu messen vorgibt. Die formale Definition: Die Validität eines Tests X ist gleich der Korrelation mit einem Kriterium Y:

Validität: $$\rho\,(X,Y) = \frac{\sigma\,(X,Y)}{\sigma\,(X) * \sigma\,(Y)}$$

Die Grundfrage psychologischer Tests bzw. ihrer Validität lautet: In wie weit kann man aus dem Testverhalten auf das reale Verhalten einer Person schliessen?
Die Validität eines Tests wird nicht gemessen, sondern aus Messungen erschlossen (Jäger, 1986, S. 272).

4.6.2. Der Validitätsbegriff

Das Validitätskonzept wird sehr inflationär verwendet. Jäger (1986) zählt folgende Begriffe auf:

> "äussere, begriffliche, curriculare, differentielle, diskriminante, divergente, ethische, faith, faktorielle, innere, intrinsische, jobanalytic, Konsens-, konvergente, logische, multiple, nomologische, ökologische, ökonomische, paramorphe, praktische, psychologische, repräsentative, soziale, strukturelle, synthetische, theoretische, trait-, Zuwachs-, ... Validität" (Jäger, 1986, S. 274).

Die APA (American Psychological Association, 1954) schlägt vor nur die folgenden Validitätsarten zu unterscheiden, da die oben genannten Begriffe als Spezifikationen dieser Validitätsarten gelten können:

a) predictive validity	(Voraussagevalidität)	a, b } Kriteriumsvalidität
b) concurrent validity	(Uebereinstimmungsvalidität)	
c) content validity	(Inhaltsvalidität)	
d) construct validity	(Konstruktvalidität)	

4.6.3. Kriteriumsvalidität

Die Voraussagevalidität und die Uebereinstimmungsvalidität schliessen auf ein ausserhalb liegendes Kriterium. Durch die Validierung wird überprüft, in welchem Ausmass der Test einen momentanen Zustand (Uebereinstimmungsvalidität) oder eine erst später beobachtbare Verhaltensweise (Vorhersagevalidität) erfasst. Bei praktischen Fragen der Klassifikation oder Selektion von Personen spielen diese Validitätsarten eine grosse Rolle. Sie werden mit der multiplen Regression berechnet. Probleme können dadurch entstehen, dass der Test nicht zuverlässig (unreliabel) ist, aber auch dadurch, dass das Kriterium nicht reliabel erfasst wurde. Bsp.: Kann durch ein Test zur Messung der Handgeschicklichkeit (MLS nach Schoppe) Leistungsunterschiede in der Genauigkeit der Bearbeitung eines Werkstückes von Lehrlingen vorhergesagt werden?

4.6.4. Inhaltsvalidität

Ist der Inhalt einer Messung repräsentativ für die Gesamtmenge des Verhaltens, das untersucht werden soll? Damit inhaltliche Validität gegeben ist, müssen die Testitems eine repräsentative Stichprobe aus der Population aller entsprechenden Testaufgaben darstellen. Man kann diese Validität nicht berechnen, sondern nur beurteilen lassen (z.B. durch Fachleute). Beim praktischen Vorgehen muss man möglichst viele Items sammeln und aus dieser Sammlung Stichproben entnehmen. Bsp.: Der MMPI zur psychiatrischen Diagnose bildet mit seinen 566 Items noch heute einen riesigen Itempool zur Entwicklung von klinischen Tests und Persönlichkeitstests ("Itemklau" ist bei der Testkonstruktion ein durchaus übliches Mittel).

4.6.5. Konstruktvalidität

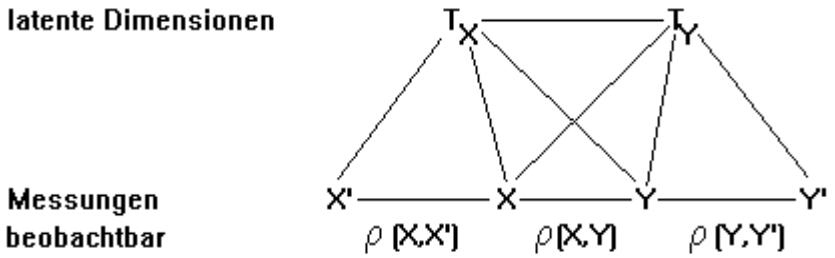
Bei der Konstruktvalidierung will man die individuellen Unterschiede zwischen den Testwerten eines Messinstruments erklären, indem man das hinter dem Gemessenen Stehende herauszufinden versucht. Es geht also um die Frage, welche psychologischen Eigenschaften oder Konstrukte die Testvarianz zustande bringen. Cronbach (1960) postuliert drei Schritte bei der Konstruktvalidierung:
a) Welche Konstrukte sind für die Testleistung verantwortlich?
b) Hypothesenbildung auf Grundlage der Theorie, in der das Konstrukt enthalten ist.
c) Empirische Ueberprüfung der Hypothesen.
Bsp.: Sarason (1958) versuchte die Validität seiner Prüfungsangstskala für Kinder (TASC) zu belegen. Dazu wurde das Konstrukt Angst als zusammengesetztes System verschiedener Angstkomponenten postuliert. Die allgemeine kindliche Angst, erfasst über eine Einschätzung der Lehrer korrelierte signifikant mit den Prüfungsangstwerten im TASC bei einer Stichprobe von 2200 Schülern (konvergente Validität). Intelligenz und Leistung, die unabhängig erfasst wurden korrelierten, wie vorausgesagt, nicht mit der Prüfungsangst (diskriminante Validität). Diese Art der

Konstruktvalidierung geht auf Campbell & Fiske (1959) zurück, die konvergente und diskriminante Validität in der "Multitrait-Multimethod Matrix" unterscheiden. Die Faktorenanalyse wird neben der multiplen Regression als Analyseverfahren für die Konstruktvalidität verwendet.

4.7. Zusammenhang zwischen Reliabilität und Validität

Durch unsere Messung bzw. Beobachtung wollen wir auf zugrundeliegende, nicht beobachtbare, latente Dimensionen schliessen. Reliabilität und Validität machen Aussagen über Zusammenhänge zwischen Beobachtungen, nicht aber zwischen latenten Dimensionen, die den Beobachtungen bzw. Messungen zugrunde liegen.
Aus dem Modell der klassischen Testtheorie kann man aber auf diese latenten Dimensionen schliessen.

Abbildung 9: Zusammenhang der beobachtbaren Messungen und den nicht beobachtbaren latenten Dimensionen



Legende: X und X' sind parallele Messungen der latenten Dimension T_X. Y und Y' sind parallele Messungen der latenten Dimension T_Y. Aus den Reliabilitäten der parallelen Messungen und der Validität zwischen X und Y kann man auf die latenten Dimensionen schliessen. Die Linien verdeutlichen die Beziehungen, die man herstellen muss.

In unserem oben genannten Beispiel ist T_X die Handgeschicklichkeit eines Lehrlings. T_Y ist das Kriterium, nämlich "exakter und sorgfältiger Umgang mit Werkstücken". T_X kann durch die MLS nach Schoppe gemessen werden (Messung X und X'). Auf T_Y können wir durch eine Einschätzung der Lehrmeister schliessen (Messung Y und Y').
Nur die Reliabilität von X, die Reliabilität von Y und die Validität von X bzgl. des Kriteriums Y sind beobachtbar. Alle Korrelationen mit den latenten Dimensionen bzw. wahren Werten sind nicht direkt beobachtbar, können aber durch Formeln erschlossen werden.

Die Verdünnungsformel (die Korrelation wird durch den Messfehler verdünnt) oder englisch "attenuation formulae" gibt an, wie hoch die Korrelation zwischen Test X und Kriterium Y höchstens werden kann, wenn die Messfehler verschwunden sind, d.h., wenn T_X und T_Y korreliert werden.

$$\rho\,(T_X,T_Y) = \frac{r(X,Y)}{\sqrt{r(X,X') * r(Y,Y')}}$$

Diese Formel gibt an wie hoch die maximale Validität eines Tests in Bezug auf ein bestimmtes Kriterium überhaupt werden kann (Idealkorrelation). Man kann so abschätzten, ob sich die Erhöhung der Reliabilität eines Tests (z.B. durch Testverlängerung) überhaupt rentiert, wenn die Korrelation der wahren Werte T_X und T_Y nicht viel höher ausfällt, als die Korrelation der Messwerte X und Y.

Durch Umformung dieser Formel kann man abschätzten, wie sich die Validität verändert, wenn man die Messung durch den Test (1.), oder die Messung des Kriteriums (2.) und schliesslich die Messung von Test und Kriterium (3.) reliabler macht.

	Validität nach Verlänge-rung		Validität vor Verlänge-rung	
1.	$\rho(X_n,Y)$	=	$\rho(X,Y) *$	$\sqrt{\frac{r(X_n,X_n')}{(r(X,X'))}}$ Reliabilität des Tests nach Verlängerung / Reliabilität des Tests vor Verlängerung
2.	$\rho(X,Y_n)$	=	$\rho(X,Y) *$	$\sqrt{\frac{r(Y_n,Y_n')}{(r(Y,Y'))}}$ Rel. des Kriteriums nach Verlängerung / Rel. des Kriteriums vor Verlängerung
3.	$\rho(X_n,Y_n)$	=	$\rho(X,Y) *$	$\sqrt{\frac{r(X_n,X_n') * r(Y_n,Y_n')}{r(X,X') * r(Y,Y')}}$

Aus der Verdünnungsformel und den Ableitungen dieser Formel ergibt sich ausserdem, dass die Reliabilität hoch sein muss, wenn die Validität zufriedenstellend sein soll.
Die Validität kann grösser sein, als die Reliabilität, aber nicht grösser als die Wurzel aus der Reliabilität.

Merke: Die Wurzel aus der Reliabilität ist die obere Schranke der Validität.

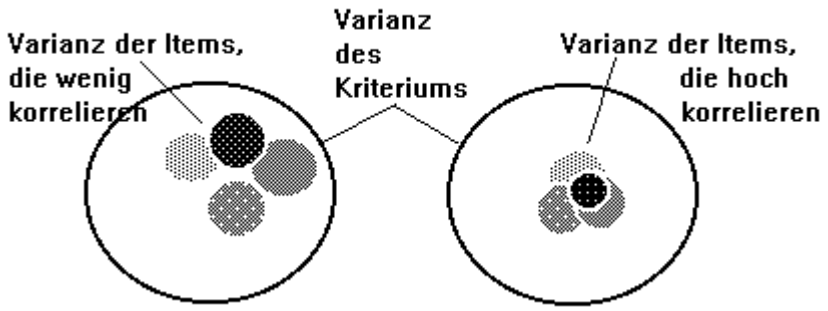
$$\rho(X,Y) \leq \sqrt{\rho(X,X')}$$

Das Attenuation Paradoxon (Solomon):
Wie sieht der genaue Zusammenhang zwischen Validität und Reliabilität aus?
Wie muss man den Test konstruieren und die Items selegieren, um bestmögliche Reliabilität und Validität zusammen zu erhalten?
Folgende Formel beschreibt die Validität als Funktion der Items x_i und x_j, die parallel sein sollen:

$$\rho\,(X,Y) = \frac{\sum_i r(x_i, Y) * s(x_i)}{\sqrt{\sum_i \sum_j s(x_i) * s(x_j) * r(x_i, x_j)}}$$

Im Nenner dieser Formel steht die Iteminterkorrelation von x_i und x_j ($\rho\ (x_i,x_j)$). Die Iteminterkorrelation sollte klein sein, wenn die Testvalidität gross ein soll. Hier tritt ein Widerspruch zwischen der Forderung nach hoher Reliabilität und hoher Validität zutage. Das Attenuation Paradoxon besagt nämlich, dass in bestimmten Fällen eine Erhöhung der Reliabilität eine Senkung der Validität bewirken kann.
Man muss sich dies vorstellen wie in Abbildung 10.

Abbildung 10: Venn-Diagramm zur Erläuterung des "Attenuation Paradoxon"



Legende: Die schwarz umrandeten Kreise stellen die Varianz des Kriteriums dar. Dies ist der Bereich, den man erfassen will. Die schraffierten Kreise entsprechen der Varianz der einzelnen Items, die verschieden grosse Bereiche des Kriteriums abdecken.

Je mehr an der Varianz des Kriteriums aufgeklärt wird, umso höher ist die Validität der Tests. Bei hoher Interkorrelation der Items (sie sind sehr homogen, d.h. sie messen dasselbe) resultiert eine hohe Reliabilität, aber nicht unbedingt eine hohe Validität.
Praktisch folgt daraus, dass man nicht Items durch immer homogenere ersetzten sollte, sondern, dass man weitere parallele Items zufügt (Testverlängerung), denn diese wirken sich nicht nachteilig auf die Validität aus.

4.8. Kritik der klassischen Testtheorie

1. Unüberprüfbarkeit der Annahmen ($X = T + F$; $E(F) = 0$; $E(T) = E(X)$; $\rho(F,T) = 0$; $\rho(F_1,T_2) = 0$; $\rho(F_1,F_2) = 0$). Dies widerspricht einem Kriterium der Wissenschaftlichkeit, nämlich der empirischen Prüfbarkeit.

2. Ein weiteres Kriterium der Wissenschaftlichkeit ist Widerspruchsfreiheit. Die im Attenuation Paradoxon gezeigt Unvereinbarkeit von optimaler Validität und Reliabilität widerspricht diesem Prinzip.

3. Korrelationskoeffizienten sind stichprobenabhängig. Da Reliabilität und Validität durch Korrelationen definiert sind gilt dies auch für diese Koeffizienten. Bsp. in Tabelle 5. Erwin und Else wurden in einer "schlechten" (Stichprobe A) und in einer "guten" Stichprobe (Stichprobe B) gemessen. Es ergeben sich unterschiedliche Mittelwerte, Standardabweichungen, Reliabilitäten, Standardmessfehler und schliesslich Konfidenzintervalle für den "wahren Wert". Bei Stichproben mit grosser Varianz (hier bei Stichprobe B) wird die Reliabilität grösser.

Tabelle 5: Stichprobenabhängigkeit von Korrelationskoeffizienten

	Stichprobe A			Stichprobe B	
Messungen	x	x'		x	x'
Erwin	90	85	Erwin	90	85
Else	85	90	Else	85	90
W.	100	105	A.	130	135
S.	105	100	J.	135	130

$\bar{x}$ =	95	95	110	110
s_X =	9,12	9,12	26,1	26,1
$r_{xx'}$ =	.80		.97	
s_F =	4,07		4,52	
Konfidenz-intervall	±8,0		±8,9	

4. Die Skalenfestlegung ist abhängig von der Itemstichprobe. Testscores auf einer Skala sagen nur im Vergleich mit anderen Probanden (Normstichprobe) etwas aus.

5. Der Messfehler wird in allen Bereichen der Skala gleich gross angenommen (Siehe auch die Berechnung der Konfidenzintervalle). Es ist aber eher anzunehmen, dass in den Extrembereichen einer Skala schlechter (oder vielleicht auch besser) gemessen wird als im mittleren Bereich. Dies würde aber eine Korrelation von Fehler und wahrem Wert nahelegen, was laut Axiom ausgeschlossen ist.

Wie ist diese grundlegende Kritik zu werten?
Fischer (1968, 1974) entwickelt eine sehr differenzierte Kritik, die aber nur zögernd angenommen wird und Hilke (1980) setzt sich ebenfalls kritisch mit der klassischen Testtheorie (und den Probabilistischen Testmodellen) auseinander.
Als Fazit bleibt, dass nicht geprüft werden kann, inwieweit die Voraussetzungen (Axiome) der klassischen Testtheorie gegeben sind. Man muss sich deshalb vergegenwärtigen dass es sich bei der KTT um ein Modell handelt, dessen Angemessenheit für den psychologischen Gegenstandsbereich zumindest hinterfragt werden muss.
Für die Praxis gilt, dass es kaum Tests gibt, die nicht auf der klassischen Testtheorie basieren (vgl. auch Kapitel 5). Die Zielsetzungen der Differenzierung zwischen Individuen und der Vorhersage auf ein aussenliegendes Kriterium scheinen jedenfalls durch die "klassisch" konstruierten Tests durchaus gegeben.

4.9. Uebungsaufgaben zur klassischen Testtheorie

1. Ein Psychologe hat eine Angstskala (A) erarbeitet. Bei der Datenerhebung lässt er gleichzeitig eine anerkannte Angstskala (B) ausfüllen. Er berechnet:

$r_{ttA} = .50$ $\quad$ $r_{ttB} = .85$

$\sigma^2(A) = 16$ $\quad$ $\sigma^2(B) = 25$

$\sigma(A,B) = 9$

a) Berechne die Validität von A

b) Wären A und B Tests, die ohne Fehler messen würden, wie hoch wäre dann die Validität?

2. Erläutere die teilweise Inkompatibilität von Reliabilität und Validität (Attenuation Paradoxon).
   a) anhand eines Beispiels
   b) anhand der relevanten Formel

3. Nenne 4 Methoden der Reliabilitätsbestimmung und deren Vor- und Nachteile.

4. Die Reliabilität wird durch Korrelationsrechnung bestimmt, welche Nachteile entstehen hierdurch?

5. Das Quadrat des Korrelationskoeffizienten r_{TX} ist :
   - ❑ die Validität von Test X
   - ❑ die Reliabilität von Test X
   - ❑ die Validität vom wahren Wert T
   - ❑ die Varianz von Test X
   - ❑ die Kovarianz von Test X

6. Ein Schulleistungstest besteht aus 30 Items und hat eine Reliabilität von .73. Der Test soll in der ganzen Schule angewandt werden. Wegen Einsparungsmassnahmen stehen dem Schulpsychologen aber nur DIN A4-Blätter zur Verfügung, auf denen gerade 10 Items untergebracht werden können. Hat es einen Sinn, den Test auf 10 Items zu reduzieren, wenn man annimmt, dass die minimale Reliabilität .50 sein muss?

7. Ein frischgebackener Testkonstrukteur sagt: "Ich habe eine Testbatterie ausgearbeitet mit über 10 Prädiktoren zur Vorhersage der Eignung für ein Psychologiestudium. Da ein Prädiktor mit .80 mit dem Kriterium "Studienerfolg" korreliert und alle anderen Prädiktoren wiederum mit diesem einen Prädiktor hoch korrelieren, ist meine Testbatterie die beste, die seit langem auf dem Markt ist. Was würdest Du ihm darauf entgegnen?

8. Ein Psychologe errechnet bei zwei parallelen Tests A und B, die die technische Begabung erfassen, folgende Werte aus:
   Test A: Mittelwert = 50 ; Varianz = 25
   Test B: Mittelwert = 50 ; Varianz = 25
   Reliabilität (Korrelation zwischen A und B) = .70
   Berechne die wahre Varianz und die Fehlervarianz von Test A.

9. Welche Methode gewährleistet die Reliabilität eines Tests zu erhöhen, ohne dass die Validität sinkt?

10. Ein Psychologe will die Reliabilität eines Persönlichkeitstests feststellen. Er gibt den Test zwei unabhängigen Gruppen von Personen (je 100 Personen). Bei der ersten Gruppe errechnet er eine Reliabilität von .60, bei der zweiten Gruppe eine Reliabilität. von .40. Wodurch kann es zu diesem Unterschied kommen, obwohl es sich um den selben Test handelt?

11. Bei der Korrelation eines Intelligenztests mit der aktuellen Schulleistung spricht man von:

    ☐ Vorhersagevalidität
    ☐ Uebereinstimmungsvalidität
    ☐ Konstruktvalidität
    ☐ Augenscheinvalidität
    ☐ nichts von dem Quatsch, sondern .....

12. Eine Testskala A aus einer Testbatterie korreliert signifikant mit einem Kriterium. Als der Untersucher sich jedoch die multiple Regressionsgleichung der Batterie zur Vorhersage des Kriteriums ansieht, muss er feststellen, dass die betreffende Skala A ein Regressionsgewicht von 0 besitzt. Woran liegt das?

13. Es gibt verschiedene Methoden der Reliabilitätsbestimmung. Trage die Namen der Methoden ein:
    a) Die Interkorrelation der Items führt zu .....
    b) Die Korrelation der Testergebnisse zwischen zwei Zeitpunkten führt

zu .....

c) Die Korrelation zweier vergleichbarer Tests, die das gleiche Merkmal messen führt zu .....

14. Wie gross ist die Reliabilität eines Tests, dessen Standardabweichung s_X=15 und dessen Standardmessfehler s_F=15 ist?

15. Untersuche die folgenden Aussagen und entscheide jeweils, ob es sich um ein Reliabilitäts-, ein Validitätsproblem oder beides handelt. Gib zudem an, um welche Art der Reliabilitäts- bzw. Validitätsbestimmung es sich handelt.

a) Ein Test wurde an der gleichen Stichprobe von Probanden zweimal durchgeführt. Der Korrelationskoeffizient zwischen den Ergebnissen beider Termine betrug r = .90.

❑ Reliabilitätsproblem ❑ Validitätsproblem ❑ beides

Bestimmungsart: _____

b) Vier Lehrer untersuchen die Items eines Tests hinsichtlich deren Relevanz für die Lehrziele im Lehrplan.

❑ Reliabilitätsproblem ❑ Validitätsproblem ❑ beides

Bestimmungsart: _____

c) Die Korrelation zwischen einem Schulleistungstest und dem Notendurchschnitt im Zeugnis beträgt r = .55

❑ Reliabilitätsproblem ❑ Validitätsproblem ❑ beides

Bestimmungsart: _____

d) Eine neu entwickelte Konservativismusskala wurde überprüft, indem sich signifikante Unterschiede zwischen Mitgliedern der Sozialdemokratischen (SP) und der Christdemokratischen Partei (CVP) ergaben.

❑ Reliabilitätsproblem ❑ Validitätsproblem ❑ beides

Bestimmungsart: _____

16. Der Korrelationskoeffizient zwischen zwei Halbtests (gemeint ist nicht die Spearman-Brown Formel) ist eine

❑ Unterschätzung
❑ exakte Schätzung
❑ Ueberschätzung

der Reliabilität des Tests.

17. Nenne mindestens zwei Faktoren, die die Validität eines Tests beeinflussen:

a) _____

b) _____

18. Welches wäre von den unten angegebenen Tests T1, T2, T3 die beste Kombination von 2 Tests zur Voraussage des Kriteriums C. Erkläre den Sachverhalt.

Korrelationsmatrix:

	C	T1	T2
T1	.43		
T2	.41	.72	
T3	.32	.04	.12

a) ❑ T1 & T2 ❑ T1 & T3 ❑ T2 & T3

b) Erklärung: _____

_____

_____

_____

19. Ein vorläufiger Wortschatztest erreichte in einer Kurzform (15 Items) eine Reliabilität von r_{tt} = .75.

a) Wieviele Items müssen mindestens hinzugefügt werden, um die Reliabilität auf .90 zu steigern?

b) Welche Gründe könnten gegen eine Testverlängerung sprechen?

20. Die Schüler Hans und Dieter haben an einem Intelligenztest (Reliabilität: r_{tt} = . 84 und Standardabweichung: $\sigma(x) = 15$) teilgenommen und sprechen danach über ihre Ergebnisse. Hans hat einen Score von 105, Dieter von 120. Dieter sagt: "Siehst Du! Das zeigt wissenschaftlich eindeutig, dass ich intelligenter bin als Du!". Kannst Du dieser Aussage aufgrund testtheoretischer Ueberlegungen zustimmen? (Grundlage des Deines Urteils sollte eine Irrtumswahrscheinlichkeit von p=.05 sein, das entspricht einem kritischen Z-Wert von Z_{Krit} = 1.96).

21. Wie gross ist die Reliabilität r_{tt} eines Tests, dessen Standardabweichung $\sigma(x) = 15$ und dessen Standardmessfehler $\sigma(F) = 15$.

- ☐ $r_{tt} = 1.00$
- ☐ $r_{tt} = .50$
- ☐ $r_{tt} = 0$
- ☐ Ueber die Reliabilität ist keine Aussage möglich