Lernzettel Testtheorie

# **Psychologische Tests**:

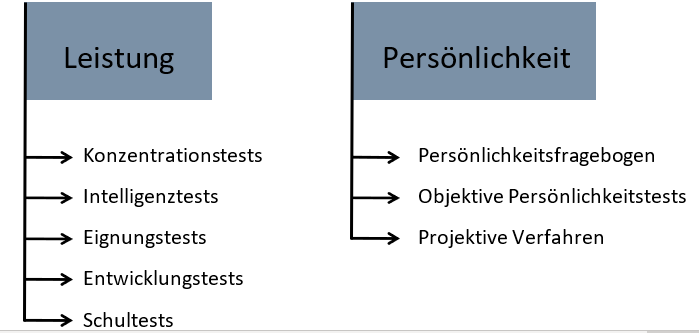
# misst in der Regel ein einzelnes psychologisches Konstrukt oder auch mehrere Konstrukte aber dann auch durch unterschiedliche Skalen -> ein Konstrukt, dass auf mehrere Items kommt

* Ziel einer quantitativen Auswertung (Zahlenwert)
* kann relativ zu Referenzgruppe interpretiert werden (75% haben höheren/niedrigeren Wert)
* Gütekriterien spielen wichtige Rolle

# **Psychotests:**

messen mehrere Konstrukte (Erfolgsorientierung, Ordnungssinn…) -> ein Item für mehrere Konstrukte

* qualitative Auswertung
* kein Vergleich mit anderen Personen
* **Tests of maximum performance** (Leistungstests) liegen immer dann vor, wenn Aufgaben entweder **richtig oder falsch** beantwortet werden können, also ein verbindlicher Beurteilungsmaßstab vorliegt.
* **Tests of typical performance** (Persönlichkeitstests) sind Tests in denen Persönlichkeitsmerkmale bzw. **Charaktermerkmale wie Eigenschaften, Motive, Interessen, Einstellungen** etc. erfasst werden.
* **-> kein richtig/falsch**



# Testtheorie

* Ziel = Merkmale erfassen die **nicht direkt beobachtbar** sind z.B. Intelligenz
* typische **beobachtbare** Verhaltensweisen für z.B. Intelligenz (hohe Allgemeinbildung, Zahlenverständnis, mentale Rotationsaufgaben…)
* auf Basis von Verhaltensindikatoren werden **Items** generiert
* **beobachtbares Lösungsverhalten** kann zu **Testscore** zusammengefasst werden -> nicht beobachtbares Konstrukt erfassen

ABER funktioniert nur wenn

1. Messung Verhaltensindikatoren objektiv (nicht von Versuchsleiter/Situation verfälscht
2. Test zuverlässig misst (immer gleiches Ergebnis bei gleicher Person)
3. verwendete Verhaltensindikatoren decken Konstrukt umfassend ab -> sonst Test nicht valide

Testtheorien stellen den theoretischen Hintergrund zur Konstruktion und Interpretation von Testverfahren dar.

Sie beschreibt die grundlegenden theoretischen Annahmen, die notwendig sind, um von einer **beobachteten, aber meist nicht messfehlerfreien Messung** auf den dahinterliegenden **„wahren Wert“** des zu erfassenden Merkmals schließen zu können.

Testtheorie stellt die **mathematischen Konzepte** bereit, um den Zusammenhang von Testverhalten und dem zu erfassenden Merkmal zu beschreiben.

# Gütekriterien

Aus den Vorgaben der Testtheorie können z.B. **Gütekriterien** und deren Berechnung abgeleitet werden

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

# Testentwurf

# Bestandteile psychologischer Tests

* Fragebogen selbst
* Instruktion
* Anweisungen zur Interpretation/Auswertung des Tests z.B. durch Profilbögen
* Manual -> Gütekriterien, Item- und Faktorenanalyse, Definition des Konstrukts, Konstruktionsschritte
* häufig knapp zusammengefasst in Artikelform

**Was** soll der Test messen?

* Definition des Konstrukts -> aus Fachliteratur oder eigene Arbeitsdefinition
* konzeptuelle Einengung oder Erweiterung des Konstrukts
* ein- oder mehrdimensional

**Für wen** soll der Test geeignet sein?

* Itemformulierungen -> Zielgruppe (Kinder, Bildungsstand, Kultur, Sprachbarrieren, bestimmte Berufsgruppen….) -> z.B. Intelligenztests für Kinder häufig in 1:1 Situation durchgeführt

**Welchen Verwendungszweck** soll der Test haben?

* Anwendungs-/Geltungsbereich Individualdiagnostik vs. Forschungszwecke (weniger hohe Ansprüche an Gütekriterien da über alle Personen aggregiert wird)
* Einsatzbedingungen -> nötige Expertise zur Durchführung
* Expertise -> v.a. bei Interaktion zwischen Versuchsleiter und Proband (paper-pencil vs. Computer, Einzel- oder Gruppensetting)

# Rationale Konstruktion -> Orientierung an Theorien

1. Konstruktdefinition auf Grundlage elaborierter Theorien -> eventuell auch Einengung/Erweiterungen möglich (Expertengruppen, Kombination verschiedener Theorien
2. Verhaltensindikatoren identifizieren
3. Items/Testentwurf entwickeln
4. Erprobung (Daten erhoben -> Itemanalyse: können Items beibehalten werden oder müssen Items entfernt oder verändert werden-> eventuell Erweiterungen…)
5. Veränderung des Itempools -> erneute Erprobung/ Revision
6. Validierung -> Misst Test das, was er messen soll?
7. Normierung = Testeichung auf gut interpretierbare Normwerte

* Beispiel I-S-T 2000R = Kombination Thurstone Primärfaktoren + fluide und kristalline Intelligenz nach Cattel und Horn + Spearmans g-Faktor (Zusammenfassung der Primärfaktoren zu einem Wert im Grundmodul)

# induktive Konstruktion

* empiriegeleitet
* wenn noch keine klaren Vorstellungen von Konstrukt vorhanden
* Itemformulierungen auf Basis vager Vorstellungen/Literatur
* exploratorische Faktorenanalyse (induktives Analyseelement) -> Bestimmung der Struktur des Konstrukts (Dimensionalität…)
* Interpretation der Faktoren
* ggf. Ableitung einer Theorie

# Beispiel -> BIG FIVE

* Sedimentationshypothese/Lexikalischer Ansatz
* Oddport, Albert: 18000 Wörter zur Beschreibung der Unterschiede menschlichen Verhaltens
* Cattell: 4500 Eigenschaftswörter im engeren Sinne
* Personen beschreiben andere Personen anhand dieser Eigenschaftswörter
* Cattell: 12 (Fremdbeurteilungs-)Faktoren
* andere Autoren ( Goldberg, John et al) nur 5 Faktoren bei gleichem Datensatz
* über verschiedene Länder und Stichproben und methodische Vorgehensweisen (Selbst-vs. Fremdbericht) konsistent

# Externale Konstruktion

* auch kriteriumsbezogene Skalenentwicklung
* Orientierung an Gruppen von Personen z.B. Kranke vs. Gesunde
* Gruppen = externes Kriterium
* Resultat -> Instrument zur Klassifikation/Diskriminierung zwischen Gruppen
* Items die besonders gut zwischen Gruppen diskriminieren
* möglichst breiter Inhaltsbereich der Items z.B. nicht nur primäre Depressionssymptome abfragen
* Höhe des Wertes kann nicht Ausmaß einer psychischen Störung gleichgesetzt werden -> höhere Werte = höhere Wahrscheinlichkeit, dass Person zu bestimmter Gruppe gehört
* orientiert sich nicht an direkt-ableitbaren Verhaltensindikatoren

# Vorgehen:

* Vorliegen einer möglichst großen Anzahl inhaltlich breit gefächerter Items
* Itemformulierungen von mehr oder weniger expliziten Hypothesen geleitet z.B. bestimmte Lebensbereiche auf die sich psychische Störung auswirkt
* Erprobung der Items an Personengruppen, die sich möglichst stark unterscheiden
* Auswahl der Items, die besonders gut zwischen Gruppen diskriminieren
* Itemauswahl besonders stark an empirischen Befunden orientiert (könnte in anderen SP abweichen)
* Notwendigkeit einer Kreuzvalidierung

+ geringe Augenscheinvalidität durch inhaltlich breit gefächerte Items (VP kann nicht auf Anhieb erkennen, was abgefragt wird, da keine direkten Symptome abgefragt werden)

+ heterogene Items

* keine dimensionale Interpretation -> Interpretation im Sinne von Wahrscheinlichkeiten zu einer bestimmten Gruppe zu gehören

Beispiel für externale Konstruktion: MMPI (Minnesota Multiphasic Personality Inventory)

* Hathaway, McKinley -> 1. Auflage 1951
* Ziel: psychische Störungen ökonomischer diagnostizieren als mit klinischen Interviews
* Breitbandverfahren
* 1000 Items (allgemeine Gesundheit, Eheliche und familiäre Beziehung, sexuelle/religiöse Einstellungen, emotionale Zustände
* Erprobung KG vs. klinisch auffälligen Personen
* 566 Items die signifikant zwischen Gruppen diskriminierten
* zu Skalen zusammengestellt (bezogen auf psychiatrische Kategorien)

# Überarbeitung -> MMPI-2

* unangemessene/nicht mehr zeitgemäße Items rausgenommen/umformuliert
* neue Items -> Inhaltsbereiche Ess- und Arbeitsstörungen
* 310 Originalversion
* 149 sprachlich überarbeitet
* 108 neu konstruiert
* 567 Items insgesamt

# Prototypen-Ansatz

Prototypische Vorstellung von Personen mit bestimmten Merkmalen ähneln sich häufig (Cantor & Mischel, 1977) -> Vorstellung Person die ordentlich… ist

-Sammlung prototypischer Vorstellungen

*Beispiel: Act-Frequency-Approach(Buss& Craik, 1983):*

Vorgehen:

-Stichprobe aus der Zielpopulation (an der Test auch später angewendet)

-Sollen an eine Person denken, die eine bestimmte Eigenschaft hat

-Prototypische Verhaltensweisen für bestimmte Eigenschaft nennen -> gesammelt -> Wie gut decken sich Verhaltensweisen?

-sehr häufig genannte Verhaltensweisen herangezogen für Ableitung von Items

-Beurteilung der Items (auf Basis einer zweiten Stichprobe) bezüglich der Prototypizität für die Eigenschaft

+ viel Wissen über Sprachgebrauch der Zielpopulation -> Wortwahl kann übernommen und in Items überführt werden => Items gut verständlich

* nicht jede Eigenschaft für Zielpopulation/Allgemeinbevölkerung zugänglich (komplexere Eigenschaften, die schwer greifbar sind z.B. Prokrastination) -> nicht alle Facetten einer Eigenschaft abgedeckt

Lösung: deduktiver Ansatz + parallel Prototypenansatz durchführen -> zusammenführen und sprachliche Informationen in deduktiven Ansatz übernehmen

# Items bestehen immer aus zwei Komponenten

* Stimulusteil/Itemstamm: Aussage, Frage, Bild, Geschichte, Rechenaufgabe…
* Reaktionsteil/Antwortformat: freie Zeile, Antwortskala, Ankreuzmöglichkeiten…

rationale/deduktive Konstruktion:

Konstruktdefinition -> Verhaltensindikatoren -> Items

# Beispiel Konstrukt Gewissenhaftigkeit:

Gewissenhaftigkeit

* zielstrebig, willensstark und entschlossen
* schulischer, akademischer und beruflicher Leistungserfolg
* übertrieben hohes Anspruchsniveau
* beinahe zwanghafte Ordentlichkeit und Arbeitssucht als Negativbeispiele
* Kontrolle von Impulsen, Wünschen, Begierden
* Selbstkontrolle bezogen auf Planung, Organisation und Ausführung von Aufgaben
* Konstrukt lässt in bestimmte **Facetten** unterteilen:

Kompetenz

Ordnungsliebe

Pflichtbewusstsein

Leistungsstreben

Selbstdisziplin

Besonnenheit

Was soll gefragt werden? -> Itemtyp

Wie soll gefragt werden? -> Itemformulierung

Wie soll geantwortet werden? -> Antwortformat

# Was soll gefragt werden?

1. Beschreibung von Reaktionen -> beobachtbares Verhalten Ich mache…
2. Eigenschaftszuschreibungen -> Ich bin…
3. Wünsche/Interessen Ich mag/würde gerne…
4. Biographische Fakten -> Ich war…
5. Einstellungen/Überzeugungen -> Man sollte…
6. Reaktionen anderer -> Andere halten mich für…

# Itemformulierung

* einfach halten
* kurz halten
* bei Unsicherheit lieber rauslassen
* jeweils nur ein Aspekt abfragen

Vermeidung von:

* langen Itemformulierungen (Schachtelsätze)
* negativ formulierte Items/doppelte Verneinungen
* Fremdwörter/Fachbegriffe
* Verständlichkeit
* Konditionalsätze
* Universalsätze/Verallgemeinerungen (selten wahrheitsgemäß)
* mehr als ein sachlicher Gedanke pro Item z.B. schnell + gerne Autofahren
* Eindeutigkeit
* Suggestion
* fehlende Passung zwischen Itemstamm und Antwortformat
* Gemeinplätze z.B. Ich würde gerne meine Ziele erreichen…

# allgemeine vs. spezifische Items

1. allgemeine Items z.B. Ich bin dominant

+ nicht viele Items benötigt

+ nicht situationsspezifisch

* Person muss sich an konkrete Situation erinnern, in der Eigenschaft verhaltensrelevant ist -> fehleranfällig (Stimmung, letzte Situation, die erlebt wurde)

1. spezifische Items z.B. Bei Gruppenreferaten an der Uni übernehme ich gerne die Führung.

+ Person muss nicht lange überlegen, da spezifische Situation vorgegeben ist

* nicht jeder muss Situation schon erlebt haben (nicht alle Personen anwendbar)
* Unterschiede zwischen einzelnen Situationen möglich
* Generierung **vieler I**tems bzgl. unterschiedlicher Lebensbereiche -> Mittelung über alle Items um allgemeines Verhalten abzufragen
* Lösung = Mischung abstrakter und allgemeiner Items

Antwortformate/Reaktionsteil

# freie Antwortformate

* Kurzaufsatzaufgaben
* Ergänzungsaufgaben
* völlig freie Antworten -> Beispiel picture-frustration-test
* schriftlich oder mündlich
* strukturiertes klinisches Interview, projektive Verfahren, Kreativitätstests

+ komplexes Denken, kreative Ideen erfassen (kein richtig/falsch

-Auswertung aufwendig -> gute Codiersysteme, Vorlagen

-eingeschränkte Auswertungsobjektivität

-Konfundierung mit verbalen Fähigkeiten/Ausdrucksfähigkeit

* eingeschränkt freie Antworten
* Wissensfragen, Rechenaufgaben..

+ Erfassung verfügbares Wissen + originelle Lösungen -> richtig oder falsch

-aufwendige Auswertung

-eingeschränkte Auswertungsobjektivität wenn Antworten komplexer

# gebundene Antwortformate

* **Ordnungsaufgaben**
* Umordnung
* Zuordnung (und Sortieraufgaben) Beispiel Würfelaufgaben IST 2000-R

+ Erfassung Wissen/Kenntnisse

+ ökonomisch, objektiv

-nur Wiedererkennen -> kein freies Abrufen

* **Beurteilungsaufgaben**
* Rating-Skala
* Analog-Skala

+ differenzierte Information

+ objektiv, ökonomisch

* unipolar vs. bipolar
* Anzahl der Antwortstufen
* Existenz einer mittleren Kategorie

+ eventuell passend

-Ausweichantwort (geringere Motivation, Verständnis…)

-> Studie Simms et al: keine Unterschiede bzgl. psychometrischer Daten (aber nur Studierende untersucht -> Generalisierbarkeit?)

* analoge vs. diskrete Skala
* analog Intervallskalenniveau erreicht aber muss bei paper-pencil ausgemessen werden
* diskret -> verbal Abstufungen größer geschätzt, meist 5-7 Stufen (Studie Rohrberger: 7 überlegen ggü. 2 Stufen aber kein weiterer Informationsgewinn bei +7 -> wenig Akzeptanz + höherer Aufwand/wenig ökonomisch
* Benennung der Kategorien
* **Auswahlaufgaben**
* Mehrfachauswahl
* MC-Aufgaben ->

+ objektiv, ökonomisch

+ Abfrage von Wissen

-kein freies Abrufen -> nur Wiedererkennen

-hohe Ratewahrscheinlichkeit

- schwierig geeignete Distraktoren zu finden

-> Distraktoren müssen disjunkt sein

* dichotome Aufgaben
* ja/nein

+ objektiv, ökonomisch

-hohe Ratewahrscheinlichkeit

-Antwort erzwungen

-wenig differenziert

* 5 Stadien bei der Beantwortung eines Items -> Fehlerquellen in jedem Stadium
* Frage verstehen
* Abrufen notwendiger Information aus dem Gedächtnis
* Urteil bilden
* Antwortwahl
* Antwortabgabe

# Frage verstehen

* Beeinflussung durch Antwortalternativen (Antwortformat)
* Itemmehrdeutigkeit -> Proband versucht aus Kontext (andere Items, Antwortmöglichkeiten) Informationen zu ziehen oder antwortet willkürlich => Item misst eventuell nicht das, was es messen soll

# Abruf notwendiger Information aus dem Gedächtnis + Urteil

* Beeinflussung durch Stimmung
* Urteil durch Beantwortung **vorangegangener Items** beeinflusst => *globales Antwortverhalten*
* kein recall and count Modell -> wird der Häufigkeit von Erleben und Verhalten nicht gerecht -> eher gut zugängliche Information aktiviert
* Beispiel Häufigkeit Fernsehen -> Einfluss durch Antwortmöglichkeiten (Gefühl mehr oder weniger als andere fernzusehen), soziale Erwünschtheit…offenes Antwortformat?
* eher ein Gefühl (nur Teilmenge der tatsächlich relevanten Situationen)
* Suche nach Maßstäben (andere Personen oder interne Vergleiche)
* Bsp. Handynutzung
* low frequency scale 20,4% gaben > 2,5h an
* high frequency scale 61,5% gaben > 2,5h an
* Einfluss durch aktuelles vs. vergangenes Ereignis + positive vs. negative Valenz -> interner Vergleich
* positiv + aktuell => Ereignis überrepräsentiert (positive Verzerrung)
* Umkehr des Effekts bei vergangenem Ereignis (Maßstab/Referenzpunkt)
* ähnliche Items sollten getrennt platziert werden
* randomisiert oder systematisch variierte Abfolge z.B. beim NEO-PI-R
* allgemeine Items sollten spezifischen Item (ähnlichen Inhalts) vorangestellt werden z.B. Lebenszufriedenheit vor Zufriedenheit mit Partnerschaft

# Antwortwahl

* Antwortmöglichkeit sollten zu Differenzierungsfähigkeit der Person passen
* Antworttendenzen
* Akquieszenz = Ja/Nein-Sage-Tendenz
* gleich viel negativ wie positiv formulierte Items
* Tendenz zur mittleren Kategorie oder zu extremen Antwortmöglichkeiten -> nimmt Varianz + falsche Merkmalsausprägung
* keine mittlere Kategorie ABER eventuell mittlere Kategorie auch passend zu Merkmalsausprägung
* wenig Antwortalternativen (wenig Differenzierung?)
* bewusste Verfälschung -> Antwortwahl ≠ abgegebene Antwort (soziale Erwünschtheit)
* Bogus Pipeline Paradigma ABER ethisch fragwürdig
* Zusicherung von Anonymität
* Soziale-Erwünschtheitsskalen -> Bereinigung in Abhängigkeit der Ausprägung auf Skala zur sozialen Erwünschtheit ABER auch fehleranfällig da eventuell auch wahrheitsgemäß beantwortet -> Überkorrektur
* indirekte Testverfahren -> schwer bewusst verfälschbar z.B. IAT

# Itemanalyse

* Bearbeitung Testentwurf durch möglichst repräsentative SP
* Kennwerte = Grundlage für Entscheidungen bzgl. der Items
* externale Testkonstruktion -> 2 Gruppen zwischen denen später diskriminiert werden soll
* Studentische SP -> leicht zugänglich, kostengünstig ABER eingeschränkte Generalisierbarkeit -> Überprüfung mit anderen SP (teilrepräsentative SP)
* Online-Panel (Mechanical Turk, Infratest) -> Teilnehmer erhalten Geld) -> repräsentativ aber hohe Kosten

# Itemschwierigkeit = Anteil der Personen, die das Item im Sinne des Merkmals beantwortet hat

* Wertebereich zwischen 0 und 1
* ACHTUNG Je höher die Schwierigkeit des Items, desto leichter ist es lösbar z.B. Schwierigkeit = 1 -> 100% haben das Item richtig beantwortet (Leistungstest)

*Wie schwer ist das Item im Sinne des Merkmals?*

* Schwierigkeit p = Anzahl aller Personen, die Item im Sinne des Merkmals beantwortet haben NR / Anzahl aller Personen, die Item beantwortet haben N **-> bei Leistungstest oder dichotomen Antwortformat**

# **Unterscheidung zwischen Powertests/Niveautests und Speed-Tests**

1. Power-/Niveautests

* so viel Zeit zur Verfügung wie nötig -> Annahme, dass alle Items bearbeitet wurden
* wenn Antwortabgabe nicht erfolgt, gilt Item als ausgelassen nicht unbearbeitet

1. Speedtests

* Item gilt als unbearbeitet ODER ausgelassen
* abhängig von Position des Items -> ausgelassen wenn mindestens ein Item danach beantwortet bzw. unbearbeitet wenn Item am Ende
* unbearbeitete Items werden in der Formel nicht berücksichtigt
* **Inangriffnahmekorrektur** da sonst bei Speedtests Itemschwierigkeit gering ausfallen würde

# Itemanalyse Vorbereitung

* negativ formulierte Items invertieren/umpolen -> können wieder in Richtung des Merkmals interpretiert werden
* Skala transformieren -> Wertebereich 0 bis… (nur Schwierigkeitsbestimmung nötig bei MW, Streuungsmaße Skalierung erhalten) VORSICHT bei vorgegebenen Mittelwert wird dieser durch Transformation verändert (siehe Übungsaufgaben)

# Formel Schwierigkeit p = MW über alle Personen/ (Anzahl der Stufen -1)

* gewünschte Itemschwierigkeit .20 - .80 -> ermöglicht Differenzierung in Extrembereichen
* extremere Schwierigkeiten können zu Varianzeinschränkung führen (aber keine strengen Grenzwerte)

wenn p = .50

* Persönlichkeitstest: Antworten schwanken um Mittelpunkt der Skala
* Leistungstest: 50% haben das Item richtig beantwortet
* Leistungstest: Schwankungen der Schwierigkeit zwischen .20 und .80 erwünscht -> Differenzierung im oberen und unteren Leistungsbereich durch sowohl Aufgaben im unteren und im oberen Schwierigkeitsbereich (nicht nur Aufteilung in zwei Lager)
* auch bei Persönlichkeitstests nicht nur mittelschwere Items erwünscht (aber nicht so zwingend wie bei Leistungstests) -> *bessere Differenzierung in Extrembereichen (z.B. zwischen Personen die überdurchschnittlich und weit überdurchschnittlich sind)*

# Itemvarianz = Differenzierungsfähigkeit eines Items hinsichtlich der untersuchten Probandenstichprobe

# Verteilungen -> überprüft mit Kolmogorov-Smirnoff-Test

* wünschenswert Normalverteilung (gute Verteilungseigenschaften)
* Schwerpunkt bei Mittelpunkt der Skala
* alle Antwortmöglichkeiten genutzt
* weniger Antworten in Extrembereichen

# Schiefewert :

* rechtssteil/linksschief (Schwierigkeitswert im oberen Bereich => negativer Schiefewert)
* linkssteil/rechtsschief (mehr Antworten im unteren Bereich) => positiver Schiefewert
* möglichst trz. alle Antwortmöglichkeiten genutzt
* kritischer Wert/Schwellenwert +-1 (oder z.T. auch +-2)

# Exzess

* spitzgipflig (positive Stauchung der Verteilung -> Konzentration auf wenige Werte) vs.
* flachgipflig (kaum Unterschiede zwischen Häufigkeiten, mit denen Antwortalternativen gewählt wurden)
* kritischer Wert/ Schwellenwert +-1 (oder z.T. auch +-2)

# Berechnung der Itemvarianz

1. für jede Person Abweichung vom MW bestimmen und quadrieren
2. Summe der quadrierten Abweichungen bilden und durch Anzahl der Personen teilen

* je stärker Personen sich unterscheiden, desto höher Varianz
* durch Quadrierung können Abweichungen aufsummiert werden (Negativzeichen entfällt)
* ohne Quadrierung wäre mittlere Abweichung/ Summenwert über alle Personen immer 0
* **maximale Varianz/Varianz bei Normalverteilung ist abhängig von Skalierung**
* Itemschwierigkeit begrenzt die mögliche Ausprägung der Varianz -> hohe/niedrige Schwierigkeit reduziert Varianz
* sollte hoch sein aber maximale Varianz wenn 50% höchsten und 50% den niedrigsten Wert angibt ABER so eine Verteilung ist nicht sinnvoll
* starke Abweichung von Normalverteilung
* nur scheinbar hohe Differenzierungsfähigkeit (Einteilung in zwei Lager -> keine Differenzierung innerhalb dieser Lager)
* Richtwert Varianz bei Normalverteilung

# Zusammenhang zwischen Itemschwierigkeit und Itemvarianz

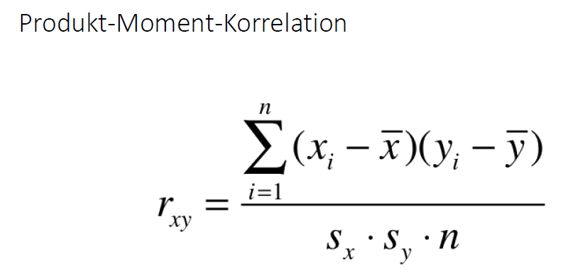
* **bei dichotomen Antwortformat:**
* Formel s^2 = p\* / (1-p)
* umgekehrt u-förmiger Zusammenhang -> maximale Varianz bei mittlerer Schwierigkeit
* **bei mehrfach gestuftem Antwortformat** ebenfalls umgekehrt u- förmiger Zusammenhang ABER maximale Varianz KANN ABER MUSS NICHT bei mittlerer Schwierigkeit erreicht werden z.B. nicht bei hohen Exzesswerten (Antworten im Extremfall nur auf Mittelpunkt der Skala zusammengeschoben -> Varianz = 0)
* mittlere Schwierigkeit kann auch anders zustande kommen z.B. wenn 50% geringste und 50% höchste Ausprägung wählen

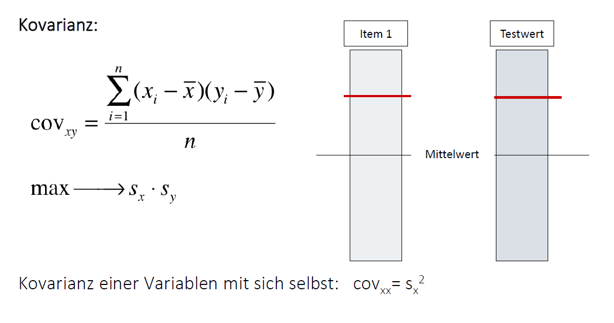
# Trennschärfe = Korrelation zwischen Itemwerten und Testwerten der Probanden (variieren gemeinsam)

* *in welchem Ausmaß, die durch das Item erfolgte Unterteilung in Löser/Nichtlöser mit der durch die Skala als Ganzes übereinstimmt*
* Kovarianz = unstandardisiert => abhängig von Skalierung (neben gemeinsamen variieren) -> Transformation auf einheitliches Maß
* maximale Kovarianz sx\*sy

-> x = Itemwert y = Testwert

* Produkt-Moment-Korrelation -> Standardisierung durch Einsetzen der maximalen Kovarianz (Produkt der beiden Streuungen) => maximaler Wert = 1
* Korrelation hoch, wenn Itemwert und Testwert gemeinsam variieren





Problem: Item für das Trennschärfe bestimmt werden soll, steckt selbst im Testwert -> Erhöhung der Trennschärfe durch Autokorrelation (Korrelation des Items mit sich selbst)

# DESHALB: Part-whole-Korrektur

* Trennschärfe wird nicht für Testwert y bestimmt, sondern Itemwert (des betrachteten Items) wird vor Bestimmung der Trennschärfe vom Testwert abgezogen (bereinigter Testwert)
* Einfluss abhängig von Länge des Tests/Anzahl der Items -> je mehr Items desto geringer Einfluss der Autokorrelation
* homogene Items -> Beitrag zur Autokorrelation geringer (weniger wichtig, ob Item in Testwert enthalten)

# Interpretation der Trennschärfe

1. nahe 1

* Konstrukt sehr inhalts- facettenarm erfasst
* Item differenziert bezogen auf das Merkmal in ähnlicher Weise wie der Gesamttest
* kaum prädiktiver Wert -> Items beziehen sich nur auf einen bestimmten Aspekt (redundant)

1. nahe 0

* Items haben wenig miteinander zu tun
* Item variiert unabhängig von anderen Items
* kein Zusammenhang zwischen Differenzierung durch das Item und Differenzierung durch den Gesamttest

1. nahe -1 -> stark negativer Zusammenhang

* Item differenziert nicht in die intendierte Richtung
* Item versehentlich nicht rekodiert? Invertiert?
* wünschenswert Trennschärfe .40 -.70

* Items inhaltlich nicht redundant
* hängen noch stark genug zusammen -> bilden dasselbe Konstrukt ab
* abhängig von Definition des Konstrukts -> bei sehr heterogenen/facettenreichen Konstrukten Trennschärfe eher geringer
* **Trennschärfe abhängig von Varianz des Items und Varianz des Testwerts**
* Varianz abhängig von Schwierigkeit des Items -> hohe/niedrige Schwierigkeit reduziert Varianz ABER Varianz essenziell für Korrelationen

# Einfluss homogener Stichproben:

* Residualvarianz = Summe der quadrierten Abweichungen der tatsächlichen Werten von den vorhergesagten Werten / Anzahl der Personen = mittlere quadrierte Abweichung der tatsächlichen von den vorhergesagten Werten
* je höher Residualvarianz im Verhältnis zur aufgeklärter Varianz, desto geringer Korrelation
* homogene SP -> Varianz der vorhergesagten Werte eingeschränkt z.B. bei ähnlichen IQ-Werte ABER Residualvarianz bleibt gleich (wenn Homoskedaszität gegeben) => Korrelation sinkt => Trennschärfe sinkt
* Vorteil heterogener SP
* WEIRD-Problematik -> homogene SP

# Abhängigkeit der Trennschärfe von:

* homogene Stichprobe -> geringere Trennschärfe
* homogenes Merkmal -> höhere Trennschärfe
* extreme Schwierigkeit der Items -> geringe Varianz -> Korrelation sinkt
* bezieht sich auf Beschaffenheit einer Skala nicht direkt eines einzelnen Items
* homogene Skalen beschreiben eng umschriebene Merkmalsaspekte
* liegt vor bei inhaltlicher und formaler Einheitlichkeit der Items einer Skala
* gute Trennschärfe .40- .70

# Richtwert >= .30 (aber in der Praxis auch geringere Trennschärfen)

# Itemanalyse/Revision

* Berücksichtigung aller Kennwerte unter Betrachtung des Einsatzzweckes
* Validität
* Reliabilität
* Itemvarianz
* Schwierigkeit
* Trennschärfe

# Fragen:

* Sind alle Antwortstufen abgedeckt?
* Ist das Item normalverteilt? (Grenzwert -> Schiefe, Exzess+-2)
* Daumenregel nach Bühner >+-1 = auffällig
* Curran, West & Finch Schiefe < 2; Exzess >7
* Miles, Shevlin <1 alles bestens
* 1> Wert < 2 Abweichung kann bedenklich sein, aber keine praktischen Implikationen
* Wert > 2 problematisch
* auch abhängig von Verteilung des Konstrukts
* Ist das Item inhaltlich valide? (Passt das Item inhaltlich zum Konstrukt?)
* Liegt die Schwierigkeit zwischen .20 und .80?
* ist die Trennschärfe ≥.30 aber < .70 ?
* Normalverteilung = wünschenswert
* symmetrisch
* normalgipflig
* bei perfekter Normalverteilung Mittelwert = Modalwert = Median

# Gründe für schiefe Verteilungen

* Merkmal in der Population grundsätzlich nicht normalverteilt
* heterogen zusammengewürfelte SP (aus Unterstichproben zusammengesetzt, in denen Merkmal grundsätzlich anders ausgeprägt ist)
* Itemauswahl nicht ausgewogen hinsichtlich der Schwierigkeit

# Lösung

* Optimierung der Itemauswahl
* Logarithmierung bei rechtsschiefer/linkssteiler Verteilung -> Werte im unteren Bereich stärker auseinander gezogen
* Flächentransformation bei rechtssteiler/linksschiefer Verteilung (unabhängig von Verteilung)

# Skalenbildung

* Faktorenanalysen untersuchen korrelative Beziehungen zwischen Variablen

# induktive Testkonstruktion -> exploratorische Faktorenanalyse

* struktursuchend -> hypothesengenerierend
* kein theoretisches Modell
* prüft, ob Konstrukt ein- oder mehrdimensional
* aus Empirie kann Modell abgeleitet werden
* latente Variable zeigt sich durch bestimmte Verhaltensindikatoren (in Form von Items ausgedrückt)
* korrelative Beziehungen zwischen Items können auf latente Variable zurückgeführt werden (keine ursächlichen Beziehungen zwischen Items) => Items variieren gemeinsam

# deduktive Testkonstruktion -> konfirmatorische Faktorenanalyse

* strukturprüfend
* Übereinstimmung zwischen empirischer und in der konfirmatorischen Testkonstruktion modellierten Datenstruktur = Hinweis auf theoretisch abgeleitete Konstruktstruktur

# Faktorenanalyse

* Charles Spearman
* g-Faktor der Intelligenz -> Korrelationen zwischen Ausprägungen in spezifischen Fähigkeiten auf Generalfaktor der Intelligenz zurückführbar
* Ziel exploratorische Faktorenanalyse : Feststellung der Anzahl latenter Variablen (Faktoren)
* Zuordnung der Items aufgrund des Einflusses der latenten Variable -> dem Faktor mit besonders starkem Einfluss zugeschrieben
* Zuordnung der Items zu bestimmten Merkmalen/Skalen muss händisch getroffen werden (nur Anzahl feststellbar)

Vorüberlegungen/Itemauswahl

* kein statistisches sondern logisch- inhaltliches Problem
* auch bei induktiver Testkonstruktion Vorüberlegungen notwendig (ähnlich wie bei deduktivem Vorgehen) ABER keine Theorie als Grundlage -> aus logisch inhaltlichen Überlegungen + empirischen Daten eigene Struktur ableiten
* Ergebnis der Faktorenanalyse abhängig von Items, die in ihr berücksichtigt werden

*Welche Faktoren/latente Variablen sollen extrahiert werden ?*

* *Auswahl von Items die diese Faktoren/Komponenten abbilden können*

# Stichprobenumfang

* Daumenregel N > 3 x Anzahl Items -> sonst geringe Stabilität/Präzision bei Messwiederholung -> je größer N desto stabiler die Schätzungen
* BESSER N> 5x Anzahl Items
* bei heterogenem Konstrukt sollte N größer sein

# Fragen:

# Welche Schätzmethode eignet sich zur Faktorisierung der Items?

* Hauptkomponentenanalyse (PCA)
* Hauptfaktorenanalyse (PFA)/Hauptachsenanalyse -> bessere >Schätzungen der Parameter als bei PCA, da geschätzter Messfehler rausgerechnet => Methode der Wahl
* Maximum-Likelihood-Faktorenanalyse (ML) -> bessere Schätzungen als PFA ABER an strengere Bedingungen geknüpft (multivariate Normalverteilung) => nur wenn auch konfirmatorische Faktorenanalyse geplant

Lassen sich die Items faktorisieren? (Zusammenhänge zwischen Item hoch genug?)

# Sind die Faktoren orthogonal/unabhängig oder oblique/abhängig?

* Welche Rotationsmethode?

# Wie viele Faktoren lassen sich extrahieren/stehen hinter der Konstruktstruktur? -> Abbruchkriterien

# Welche Items lassen sich welchem Faktor zuordnen?

* Doppelladungen -> starke Korrelation eines Items mit mehreren Faktoren => Fürntratt-Kriterium
* Nullladung -> keine substanzieller Zusammenhang mit einem Faktor
* Itemgruppen, die untereinander stark korrelieren lassen sich zu Faktoren zusammenfassen (ausreichend Korrelationen zwischen Items vorhanden?)
* bei vielen Items Indizes um Faktorisierbarkeit statistisch zu überprüfen:

# Bartletts Signifikanztest

* Vergleich empirische Korrelationsmatrize mit Einheits-/Nullmatrize (alle Interkorrelationen = 0)
* bei starker Abweichung empirischer Korrelationsmatrize von Einheitsmatrize -> ausreichend Interkorrelationen zwischen Items vorhanden

# overall-Test

* Test auf signifikante Korrelationen über alle Items hinweg ABER Signifikanz auch durch starke Korrelation zwischen Zweiergruppen (ungünstig da Datenreduktion nur 50%) oder starke Korrelationen nur zwischen einem Teil der Items

# KMO (Kaiser-Meyer-Olkin-Koeffizient)

* abhängig von Höhe der Partialkorrelationen zwischen Items -> unter Auspartialisierung aller anderen Items
* Partialkorrelation = Korrelation zwischen Regressionsresiduen *-> Wie hoch korrelieren Items, wenn alle anderen Variablen konstant gehalten werden? bzw. Wie sieht der Zusammenhang zwischen Items aus, wenn um Einfluss einer Drittvariable bereinigt?*
* Gibt es einen substanziellen Zusammenhang zwischen Variablen oder kann dieser auf eine Drittvariable zurückgeführt werden?
* Residualvarianz = nicht aufgeklärte Varianz durch z

= Gesamtvarianz – Korrelation zwischen z und y – Korrelation zwischen z und x

* Korrelation zwischen x und y nach der Auspartialisierung von z entspricht Partialkorrelation => Korrelation zwischen x und y reduziert sich im Vergleich zu vor der Auspartialisierung
* je mehr Items pro Skala , desto größer Wahrscheinlichkeit, dass Items zusammen viel überlappende Varianz ausschneiden und somit Partialkorrelation verringern
* Formel:
* Zähler = Summe aller quadrierten Interkorrelationen zwischen Items
* Nenner = Summe aller quadrierten Interkorrelationen zwischen Items + Summe aller quadrierten Partialkorrelationen (sollte möglichst klein ausfallen -> wenn 0 dann KMO =1 bzw. wenn Summe Partialkorrelationen = Summe Interkorrelationen -> KMO = .50)
* Interpretation:
* .50 - .59 schlecht
* .60 - .69 mäßig
* .70- .79 mittel
* .80- 89 gut
* > .90 sehr gut
* kritischer Wert .50 -> wenn < sind Items nicht faktorisierbar
* kritischer Wert unterschritten wenn Item 1 nur mit 2 bzw. 3 nur mit 4 und 2 korreliert -> KMO-Wert erhöht (< .50) => Datenreduktion kaum möglich
* Partialkorrelation sogar noch größer als Ursprungskorrelation
* Semipartialkorrelation: z nur aus einer der beiden Variablen rauspartialisiert

Faktorwert = gewichteter Summenwert über alle Items

* jede Person erhält zusätzlich zu Itemwerten einen Faktorwert (für jeden Faktor eigener Faktorwert)
* Faktorwert = z-standardisierte Werte (Personen mit durchschnittlicher Ausprägung auf dem Faktor Wert = 0)
* (M = 0.00; SD = 1.00; Range: –∞ bis + ∞ bzw. -3.00 bis 3.00 99.74%)
* Items gehen gewichtet in Faktorwert ein (abhängig von Stärke des Zusammenhangs zwischen Items und Faktor) -> werden nicht nur aufaddiert
* Faktoren werden nacheinander/sukzessive extrahiert
* Faktoren erklären sukzessive maximale Varianz
* Faktoren sind wechselseitig voneinander unabhängig -> Prinzip der maximalen Varianzaufklärung (sonst würden alle Faktoren zu selben Lösung führen -> erster Faktor nimmt bereits die Position ein, die zur maximalen Varianzaufklärung führt wenn Unabhängigkeit aufgehoben, würde sich der zweite Faktor auf den ersten Faktor legen)

# Darstellung im Personenraum

* Punkt = Variable -> Position der Punkte geben Werte der Personen wieder
* Achsen = Personen
* je näher Punkte beieinander, desto höher Korrelation zwischen den Items
* z-Werte -> überdurchschnittliche Werte immer durch unterdurchschnittliche Werte ausgeglichen
* 1. Faktor -> maximale Varianz/ Summe der Ladungsquadrate erreicht bei Position im Raum mit größtmöglicher Nähe zu Variablenpunkten (eigentlich ebenfalls Punkt im Raum)
* 1. Faktor mit größter Varianzaufklärung -> Extrahierung des Faktors 1 aus der **Gesamtvarianz**
* 2. Faktor muss orthogonal zu Faktor 1 liegen (Vektoren müssen 90-Grad-Winkel aufspannen)
* kann sich nur begrenzt im Raum bewegen -> kann nur maximale Varianzaufklärung aus der **Restvarianz** erreichen
* 3. Faktor wiederrum nur maximale Varianzaufklärung aus Restvarianz (die Faktor 1 und 2 übrig gelassen haben)
* kann bei 3 Personen nur noch eine Position im Raum einnehmen (eigentlich mindestens so viele VP wie Items)
* Items sehr heterogen (weit auseinander) -> mehrere Variablennester => Hinweis auf mehrere Faktoren
* Faktor 1 klärt noch nicht die gesamte Varianz auf
* noch Potenzial für Faktor 2 übrig
* Items sehr homogen/nah beieinander
* Faktor 1 klärt bereits fast die gesamte Varianz auf
* Faktor 2 aus statistischer Sicht nicht mehr relevant
* Punkte liegen genau aufeinander
* r = 1 -> Faktor 1 klärt 100% Varianz auf
* PCA -> grundsätzlich so viele Faktoren wie Items extrahierbar möglich ABER widerspricht dem Ziel der Datenreduktion
* finale Lösung eher weniger Faktoren (abhängig von Korrelationsmuster)

quadrierte Korrelationsmatrix -> quadrierte Korrelationen zwischen Faktorwerten und Variablen/Itemwerten

Faktorladung a

* Varianzaufklärung, die der Faktor an der Variablen leistet (nicht umgekehrt -> Korrelation zwischen Werten auf Item und Faktorwert) => Faktor produziert Varianz in den Items
* je näher Item an Faktor, desto höher Korrelation -> desto höher Faktorladung
* Faktorladung und Faktorwert untrennbar miteinander verbunden -> verändern sich gemeinsam
* Position des Faktors wird nach und nach verschoben bis optimale Position erreicht
* Formel Faktorwert = a1\*Item 1 + a2 \* Item2…

Fundamentaltheorem/Bestimmungsgleichung

Itemwerte/Rohwerte einer Person bestimmen

Formel:

x (Itemwert) = Faktorwert \* Faktorladung

# **Kommunalität h^2 = Varianzaufklärung, die alle extrahierten Faktoren (latenten Variablen) zusammen an einer Variablen (Item) leisten**

* Summe der quadrierten Korrelationen zwischen (Faktoren und dem Item) über alle Faktoren hinweg
* Varianzaufklärung die alle latenten Variablen (Faktoren) zusammen an einem Item leisten
* Unabhängigkeit der Faktoren erlaubt Aufsummierung (klären unabhängig Varianz in Item auf)
* Kommunalität wird größer wenn zusätzliche Faktoren hinzugenommen (zusätzliche Varianzaufklärung)

# **Eigenwert = Varianzaufklärung die ein Faktor/ latente Variable an allen Variablen leistet**

* Summe der quadrierten Korrelationen zwischen Faktor und Items (Faktorladungen) für einen Faktor über alle Variablen hinweg
* Items gehen z-standardisiert in Analyse ein (SD = 1 Var = 1) -> maximaler Eigenwert entspricht Anzahl der Items = 100 % Varianzaufklärung/gleiche Position im Raum r= 1)
* Formel zur Berechnung: Wie viel Prozent klärt ein Faktor an allen Items auf ?:

Eigenwert (des Faktors) / Anzahl der Items (=maximaler Eigenwert) \* 100 = Varianzaufklärung durch den Faktor in %

# **Faktorwert = gewichteter Wert (gewichtete Summe aller Items), der den Ausprägungsgrad einer Person auf einem Faktor darstellt**

* Gewichtung der Items erfolgt anhand der Faktorladungen

= a1 \* Item 1 + a2 \* Item 2…

* a = Faktorladung (Korrelation zwischen Faktor und Item)
* Faktorwert = z-standardisierte Werte (Personen mit durchschnittlicher Ausprägung auf dem Faktor Wert = 0)

(M = 0.00; SD = 1.00; Range: –∞ bis + ∞ bzw. -3.00 bis 3.00 99.74%)

* Faktorwerte abhängig von Faktorladungen und umgekehrt => untrennbar miteinander verbunden (verändert sich immer gemeinsam)
* Position des Faktors wird nach und nach verschoben (Position die zur maximalen Varianzaufklärung führt)

# Festlegung Anzahl Faktoren/Abbruchkriterien

* Kaiser-Guttmann
* Screeplot
* Paralellelanalyse
* basieren auf Betrachtung der Eigenwerte (Varianzaufklärung eines Faktors über alle Variablen hinweg)
* je später ein Faktor extrahiert wurde, desto kleiner sein Eigenwert
* geringer Eigenwert wenn Faktor kaum mit Variablen korreliert
* theoretische Festlegung

# Kaiser-Guttmann-Kriterium

* Extrahierung aller Faktoren mit Eigenwert > 1

ABER Faktoren kommen recht schnell auf Eigenwert> 1 (je mehr Items berücksichtigt werden desto mehr Faktorladungen fließen in Eigenwert ein -> desto größer Eigenwert)

* viele inhaltlich nicht interpretierbare (nicht relevante ) Faktoren extrahiert
* ungünstig v .a. wenn eindimensionales Konstruktstruktur angestrebt
* abhängig von Beschaffenheit des Konstrukts -> Kaiser-Guttmann-Kriterium eher bei heterogenen/komplexeren Konstrukten

# Screeplot

* abfallender Eigenwerteverlauf -> 1. Faktor mit höchstem Eigenwert
* Knick/Schwelle ab der Eigenwerte gleichbleibend niedrig -> Extrahierung der Faktoren die vor Knick liegen
* wenn nicht eindeutig -> Kombination mit Parallelanalyse

# Parallelanalyse

* Welche Eigenwerte sind signifikant? Welche Faktoren klären signifikant Varianz auf?
* auch nutzbar um eindeutigen Screeplot abzusichern
* zweiter Eigenwerteverlauf => **Zufallsverlauf** unter der Annahme dass **Items unkorreliert** (wahre Eigenwerte = 0)
* Stichprobenfehler: je kleiner n desto wahrscheinlicher sind Zufallskorrelationen zu beobachten (eigentlich Nullkorrelationen) -> Eigenwerteverläufe die deutlich von 0 abweichen
* Konvention 1000-3000 SP des Umfangs n z.B. n =150
* Zufallseigenwerteverlauf durch Mittelung über alle 1000 Zufalls- SP (jeweils 50% der Zufallseigenwerte die über oder unter Eigenwerteverlauf liegen)

ABER es reicht nicht aus zu schauen ob empirischer Eigenwert über Zufallseigenwerteverlauf liegen -> immer noch 50% Wahrscheinlichkeit dass H0 zutrifft

* angestrebt ist Wahrscheinlichkeit der **empirischen Daten** unter der **Bedingung der H0 < 5%** => signifikant
* Betrachtung **95%-Perzentil** => Ermittlung des Eigenwerts für jeden Faktor der auf Perzentil liegt
* **empirische Eigenwerte** die **über 95%-Perzentillinie** liegen treten unter der **Bedingung der Nullhypothese** mit einer Wahrscheinlichkeit **< 5%** auf

# Rotation

* bei sehr homogenen Konstrukten eventuell bereits ein Faktor ausreichend
* oder heterogene Konstrukte -> mehrere Faktoren (nach Extraktion des 1. Faktors noch genügend Restvarianz übrig)
* Problem Doppelladungen der Items (Korrelation mit mehreren Faktoren (lassen sich so nicht interpretieren)
* Veränderung der Position (Rotation) der Faktoren in jeweils eine andere Itemwolke
* Ziel: Einfachstruktur -> manche Items sollten stark andere gering mit dem Faktor zusammenhängen/korrelieren
* Items lassen sich Faktoren zuordnen
* orthogonale Rotation -> Faktoren stehen rechtwinklig zueinander
* oblique Rotation -> Faktoren nicht rechtwinklig zueinander
* Korrelationen zwischen Faktoren zugelassen (Ausmaß kann im Statistik-Programm graduell festgelegt werden) => sinnvoll wenn Variablengruppen nah beieinander
* analytische Rotation
* graphische Rotation

# Varimax-Rotation

* Ziel = Varianz maximieren
* bezieht sich **Varianz der Ladungsquadrate (quadrierte Faktorladungen) pro Faktor** (einige Items sollen (nach der Rotation) hohe andere geringe Zusammenhänge mit dem Faktor aufweisen)

Beispiel: Itemkorrelationen mit dem Faktor nach Extraktion der Faktoren zwischen .4 und .5 (geringe Varianz) -> nach Rotation Korrelationen der Items mit dem Faktor zwischen .1 und .8 (höhere Varianz)

=> günstiger für die Zuschreibung und Interpretation der Variablen, wenn einige Items hoch und andere gering mit dem Faktor korrelieren

=> Kommunalität und Gesamt-Varianzaufklärung durch Rotation unverändert ABER Verteilung der Varianzaufklärung verändert sich *(Anteil der Varianz die ein einzelner Faktor aufklärt, kann sich durch Rotation verändern aber nicht der Varianzanteil, den alle extrahierten Faktoren zusammen aufklären)*

*Formel für Varianzaufklärung mehrere Faktoren zusammen*

*Summe der Eigenwerte / Anzahl Items (maximale Varianzaufklärung) \* 100 = Varianzaufklärung durch Faktoren in %*

# Itemauswahl aufgrund der Ladung

IaI < .30 Item ist vernachlässigbar

IaI ≥ .40 Item kommt für den Faktor in Frage

IaI ≥ .60 Item sollte berücksichtigt werden

* a = Faktorladung (Korrelation Item mit dem Faktor)

# Fürntratt-Kriterium

Problem bei Doppelladungen eines Items (auf mehrere Faktoren)

* Fürntratt-Kriterium: a^2/h^2 > .50 -> Anteil der des Faktors an der Kommunalität des Items größer 50% (Anteil an der Kommunalität des größeren Faktors groß genug?)
* problematisch wenn nur 2 Faktoren extrahiert
* erweitertes Fürntratt-Kriterium
* Anteil des größeren Faktors an der Kommunalität eines Items muss um 25% größer sein als Anteil des zweiten Faktors
* Anteil an Kommunalität des 1. Faktors >62,5% (vs. 2. Faktor < 37,5%)
* *Items die nach Rotation noch Doppelladungen oder Nullladungen aufweisen müssen überarbeitet oder entfernt werden*
* Markiervariablen (besonders typische Items für latente Variable) zur Interpretation der Faktoren/Faktorenstruktur -> bereits bei Itemgenerierung festgelegt

# KTT (Klassische Testtheorie)

* älteste Messtheorie (Gulliksen 1950) = Messtheorie/Reliabilitätstheorie
* Axiome/Grundannahmen zur Beschaffenheit von Messungen
* liefert theoretische Begründung der Reliabilität
* Axiome bilden Basis zur mathematischen Ableitung der Reliabilität

**zentrale Annahme: Messungen sind fehlerbehaftet => Fehlerstreuung durch Messfehler**

* Test kann aber muss nicht unbedingt den wahren Wert abbilden

# unsystematisch

* Testkonstruktion z.B. durch mehrdeutige Items -> klar formulierte Items (keine doppelten Verneinungen, Fremdwörter…nur ein Aspekt pro Item)
* Testdurchführung -> Motivation, Testsituation, Auftreten VL-Leiter => standardisierte Durchführungsbedingungen
* Testauswertung -> Fehler bei Bestimmung des Testwerts => Schablonen

# systematisch (nicht berücksichtigt) z.B. Waage immer + 1 Kilo -> Rangreihe ändert sich nicht -> Interpretierbarkeit relativ zur Referenzpopulation gegeben

* möglichst bei Testkonstruktion berücksichtigt
* schwerer zu identifizieren (treten auch bei Messwiederholung auf)
* meist beide Fehlerarten kombiniert
* Antworttendenzen (systematisch) -> können zwischen Personen unterschiedlich stark auftreten -> invertierte Items
* Übungseffekte

**X (Testwert) = T (wahrer Wert) + E (Messfehler)**

* E schwankt mal positiv/mal negativ um wahren Wert

Axiome

# Existenzaxiom

Erwartungswert einer Messung = wahrer Wert einer Person

* Mittelwert über viele Messungen hinweg

ABER Messungen können bei Messwiederholung auch um systematisch nach oben/unten verzerrten Wert schwanken

* Über- und Unterschätzungen sind gleich wahrscheinlich
* kleine Messfehler sind wahrscheinlicher als große => Glockenkurve/Normalverteilung (Messfehler legt sich gleichmäßig um T)

# Verknüpfungsaxiom

* beobachteter Wert = wahrer Wert + Messfehler
* X = T + E
* aus Axiom 1 ergibt sich, dass der Erwartungswert des Messfehlers (über viele Messungen hinweg) Null ist
* E(X) = 0

# Unabhängigkeitsaxiom

* Korrelation zwischen den Messfehlern E und den wahren Werten T eines Tests sind unabhängig voneinander
* Messfehler weisen unabhängig vom wahren Wert immer gleiche Verteilung um T auf
* Test misst im unteren genauso gut wie im oberen Wertebereich
* Corr (T;E) = 0

weitere Annahmen (Zusatzannahmen)

* Messfehler zweier Tests sind unabhängig voneinander

Corr (E(A),E(B)) = 0

* Messfehler in Test A sind unabhängig von wahrem Werten in Test B

Corr (E(A), T(B)) = 0

**Reliabilität =** Anteil der Varianz der wahren Werte T (Varianz zwischen Personen) an Varianz der beobachteten Werte X (Varianz der tatsächlich gemessenen Werte innerhalb einer Population)

**= Genauigkeit mit der ein Merkmal erfasst wird**

* größerer Messfehler => beobachtete Varianz größer -> Anteil der wahren Varianz an der beobachteten Varianz sinkt
* Reliabilität sinkt
* Rel. = Var(T) / Var (X)
* homogene SP (z.B. nur Psychologiestudierende -> homogen bzgl. IQ)
* Reliabilität geringer da Messfehlervarianz gleich hoch aber Varianz der wahren Werte geringer

# Ableitung der Reliabilität

* mehrfache Wiederholung der Gewichtemessung pro Person
* theoretisch Mittelwert bei Testwiederholungen = wahrer Wert -> Erw (X) = T
* Varianz der Mittelwerte = wahre Varianz

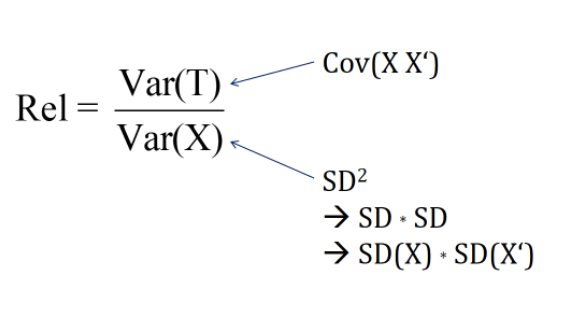
ABER unendlich viele unabhängige Messungen in der Praxis zu aufwendig + Erinnerungseffekte

* Test (für den Reliabilität bestimmt werden soll) bei gleichen Personen unter identischen Bedingungen zweimal durchgeführt
* Cov (X, X´) = Cov (T,T`) + Cov (T, E´) + Cov (T´,E) + Cov (E, E´) (alle Kombinationen von Kovarianzen zwischen wahren Werten und Messfehlern)
* X = T+E
* X´ = T´ + E´
* Cov (T, E) = 0
* Cov (T,E´) = 0
* Cov (T´,E) = 0
* Unabhängigkeitsaxiom: Messfehler hängen weder mit wahren Werten noch mit anderen Messfehlern zusammen (Teile des Terms, die sich auf Messfehler beziehen können 0 gesetzt werden)
* Cov (X,X´) = Cov (T, T´)

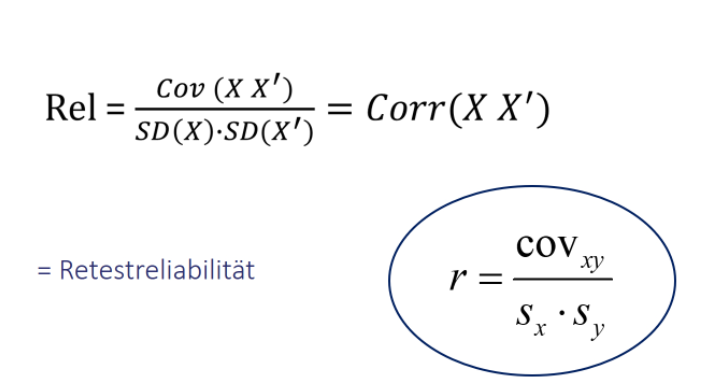
Retest-Reliabilität direkt aus der KTT ableitbar

Rel = Var (T)/ Var (X)

* Cov (X,X´) = Cov (T,T´) = Var (T)
* Var (X) = SD (X) \*SD (X´) = Varianz der beobachteten Werte
* Rel = Cov (X, X´) / SD (X) \* SD (X´) = Corr (X, X´) = Korrelation zwischen den Wertereihen über alle Personen



* Varianz der wahren Werte lässt sich bestimmen durch Kovarianz der beobachteten Werte beider Messzeitpunkte
* Varianz der beobachteten Werte durch zweifache Messung nicht quadrierte Streuung t1 oder t2 sondern Produkt der Streuung der beobachteten Werte beider Messzeitpunkte -> beobachtete Streuungen beider Messzeitpunkte berücksichtigt



* Einsetzung in Formel zur Berechnung der Reliabilität
* um Reliabilität im Sinne der KTT zu bestimmen werden zwei Messungen desselben Testes pro Person + Korrelation dieser beiden Wertereihen über alle Personen benötigt

# Interpretation der Reliabilität

r = .80

* 80% der beobachteten Varianz auf Varianz der wahren Werte zurückzuführen
* 20% der beobachteten Varianz auf Fehlervarianz zurückzuführen

# Vorteile und Grenzen der KKT

+ sparsame Theorie, die mit wenigen Grundannahmen auskommt

+ praktische Ableitungen zur Testkonstruktion, Reliabilitätsschätzung und darauf bezogene Minderungskorrektur, Konfidenzintervalle

* Axiome empirisch nicht prüfbar und nicht durchweg plausibel
* Unabhängigkeit des Messfehlers vom wahren Wert nicht unbedingt zutreffend (Intelligenztests messen in extremen Bereichen eventuell weniger genau)
* Messfehler verteilt sich nicht unbedingt gleichmäßig um wahren Wert
* Annahme eines invarianten wahren Werts (könnte sich auch zwischen Messungen verändern)
* Parameter der KTT sind populations-/stichprobenabhängig => schwankende Reliabilitäten in unterschiedlichen Substichproben (z.B. geringere Reliabilität in homogenen Stichproben) \_> Reliabilität eventuell altersabhängig
* systematische Messfehler gänzlich unberücksichtigt
* Skalenniveau oft nicht beachtet
* eigentlich Intervallskalenniveau als Voraussetzung aber viele Antwortformate können dies nicht erreichen

# Reliabilität

= Präzision/Genauigkeit mit der ein Test ein Merkmal erfasst

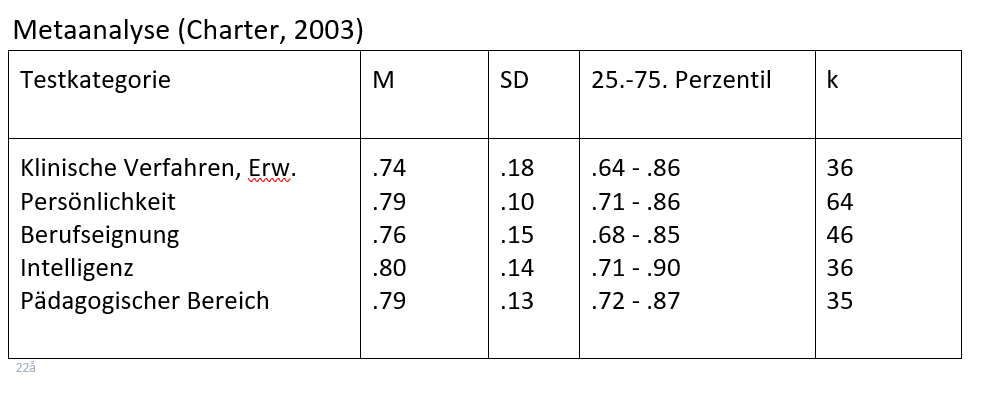
* ist unabhängig davon, ob der Test misst, was er messen soll (unabhängig von der Validität des Tests)

KTT -> Reliabilität entspricht dem Varianzanteil der wahren Werten an der Varianz der beobachteten Werte

* Rel. = Var(T) / Var(X)
* aus der KTT bzw. ihren Axiomen ist die Bestimmung der Retest-Reliabilität ableitbar

# Retest-Reliabilität (auch Stabilitätsmaß genannt)

* 1 SP
* 1 Test
* 2 Durchführungen -> jede Person erhält zwei Testwerte
* Korrelation der Testwerte aus Testdurchführung mit Testwerten aus Testdurchführung 2
* abhängig von Zeitintervall zwischen den Durchführungen
* Anwendung eher bei zeitlich stabilen Merkmalen z.B. Persönlichkeit, Intelligenz



* kein Zeitintervall angegeben
* kurzes Zeitintervall
* Transfereffekte wahrscheinlicher -> Erinnerungs-,Lern und Übungseffekte (Wie lange bleiben Transfereffekte bestehen?)
* Reliabilität wird überschätzt
* längeres Zeitintervall
* auch relativ stabile Merkmale z.B. Persönlichkeitseigenschaften können sich verändern (z.B. durch kritische Lebensereignisse…)
* personenspezifische Veränderungen auch durch neurologische Ausfälle, medizinischen Krankheitsfaktor, Medikamenteneinnahme…
* Reliabilität wird unterschätzt

*ABER KTT nimmt an, dass wahrer Wert stabil bleibt*

* Länge des Zeitintervalls Resultat praktischer Überlegungen (Intention des Tests)
* Beispiel IQ-Test zur Vorhersage des Berufserfolgs => mindestens 2 Jahre-Intervall da viel späteres Ereignis vorhergesagt werden soll
* häufig verschiedene Messabstände gewählt (nach 1 Monat, 2 Monaten, 6 Monate…)

# Paralleltest-Reliabilität

* 1 SP
* je 1 Durchführung (2 Durchführungen)
* 2 Tests
* Jede Person führt Test A und Test B durch
* Tests müssen parallel/äquivalent sein -> Paralleltest muss sich auf dieselben wahren Werte beziehen *=> Prüfung mit Strukturgleichungsmodellen*
* Erinnerungseffekte umgehen -> andere Fragen bei Test B => kein Transfer (Antworten können nicht übertragen werden
* Messfehler (Messgenauigkeit) Test A = Messfehler Test (B)
* Messfehler der Tests müssen unkorreliert sein
* Korrelation der wahren Werte soll r=1 sein
* nur in bestimmten Bereichen eingesetzt -> Leistungstests (Intelligenztest…) => nicht abschreiben können
* mit hohem Aufwand verbunden
* Problem: Übungseffekte nicht aufgehoben

DESHALB cross-over-Design

Gruppe 1 : erst Test A -> dann Test B

Gruppe 2: erst Test B -> dann Test A

* Übungseffekte auf beide Tests gleich verteilt
* Mittelwerte und Streuungen gleichen sich wieder an, wenn Tests äquivalent

# Testhalbierung

* 1 SP
* 1 Test
* 1 Durchführung
* Korrelation zweier (paralleler/äquivalenter) Testhälften
* MW, Streuungen der Testhälften müssen identisch (Feststellung durch Strukturgleichungsmodelle) sein SONST wird Reliabilität unterschätzt
* keine Parallelität der Testhälften häufig bei heterogenen Tests
* Odd-Even Methode (gerade/ungerade)
* Halbierung nach laufender Nummer
* Zufallsaufteilung
* bei Speedtests nach Zeit
* Itemzwillinge (Trennschärfe, Inhalt, Schwierigkeit) => sicherste Methode (Test sollen möglichst äquivalent sein)

Problem: Korrelation der Testhälften unterschätzt die Reliabilität des Gesamttests

* Korrektur durch Spearman-Brown-Formel

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* je länger ein Test desto größer die Interkorrelationen zwischen den Items

ABER

+ nur eine Durchführung -> höhere Ökonomie, Zumutbarkeit

+ auch bei zeitlich instabilen Merkmalen z.B. aktuelle Stimmung

*+ eventuell Vorteil zu interne Konsistenz -> Herstellung parallele Testhälften durch Itemzwillinge*

# Interne Konsistenz

* 1 SP
* 1 Test
* 1 Durchführung
* Test wird in so viele Untertest unterteilt, wie Items vorhanden (jedes Item wie eigner Test behandelt)
* Cronbachs α schätzt die durchschnittliche Interkorrelation aller Testitems
* nach oben korrigiert durch Spearman-Brown-Formel:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* ohne m würde die durchschnittliche Interkorrelation zwischen den Item die Reliabilität unterschätzen
* nur sinnvoll bei homogenen Tests => Parallelität zwischen Items besteht nicht/kann nicht hergestellt werden
* bei Merkmalen, bei denen lediglich einmalige Erfassung sinnvoll (Befindlichkeitsfragebögen) -> zeitlich nicht-stabile Merkmale
* Interne Konsistenz standardmäßig immer bestimmt (einfach zu bestimmen)

ABER Beachtung homogener vs. heterogener Test

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

# Wie hoch sollte die Reliabilität sein?

* so hoch wie möglich aber keine allgemeingültige Aussage möglich, da Reliabilität abhängig von mehreren Bedingungen
* Art des Merkmals (Einstellung vs. Intelligenz)
* Einsatzbereich (Individual- vs. Kollektivdiagnostik)

gering < .80

mittel .80 - .90

hoch > .90

# Beispiel für Richtlinien zur Bewertung der Reliabilität

* für Retest-, Paralleltest, interne Konsistenz und Interrater-Reliabilität

Niveau 1: Tests für wichtige Entscheidungen auf individueller Ebene z.B. Personalauswahl

ungenügend < .80

ausreichend .80 - .90

gut > .90

Niveau 2: Tests für weniger bedeutsame Entscheidungen auf individueller Ebene z.B. Fortschrittskontrolle

* ungenügend < .70
* ausreichend .70 -.80
* gut > .80

Niveau 3: Tests für Untersuchung auf Gruppenniveau

ungenügend < .60

ausreichend .60- .70

gut > .70

# Einflussfaktoren auf die Reliabilität

* Homogenität/Heterogenität der Testitems (Trennschärfe)
* Varianz und Verteilungsmerkmale (Schiefe, Exzess) der Testwerte
* Messfehler
* unsystematische/ vorübergehende Messfehler => situationsbedingtes Antwortverhalten
* systematische Messfehler (Antworttendenzen, Übungseffekte)

# mögliches Vorgehen

1. Reduktion unsystematischer Messfehler

* standardisierte Testbedingungen
* klar formulierte Items
* klare Instruktionen

1. Erhöhung der Homogenität der Items

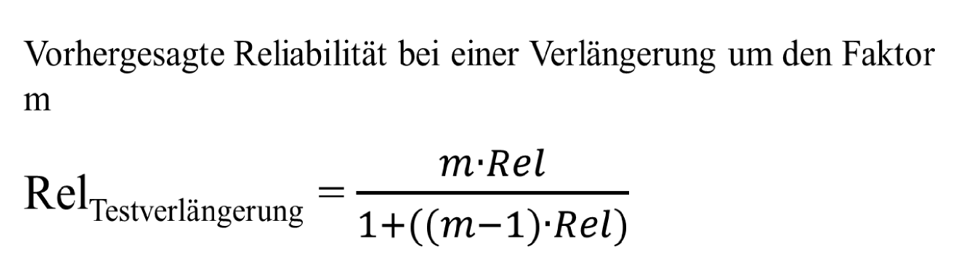
* Entfernung der Items mit geringer Trennschärfe

= Alpha-Maximierung (Maximierung Cronbach´s Alpha)

ABER Gefahr von inhalts-/facettenarmen Merkmalen -> geht auf Kosten der Inhaltsvalidität

1. Testverlängerung

* erhöht den Anteil der wahren Varianz an der beobachteten Varianz
* Spearman-Brown-Formel:



# Faktoren die Korrelation zwischen Tests beeinflussen

* Reliabilität
* Homogenität SP
* Varianz
* Validität
* Konstruktnähe

# Standardmessfehler : gibt an wie stark die Messfehler um den/die wahre(n) Wert(e) der Person(en) streuen

= die geschätzte Streuung der beobachteten Messungen einer Person bei wiederholter Messung

* geschätzte Streuung ist abhängig von der [Reliabilität](https://lms.uni-mainz.de/moodle/mod/resource/view.php?id=981829) des Testes (je höher die [Reliabilität](https://lms.uni-mainz.de/moodle/mod/resource/view.php?id=981829), desto kleiner der geschätzte Standardmessfehler) sowie der Streuung sx  des Merkmales (der Testwerte) in der Population ab (je größer die Streuung der Testwerte in der Population, desto größer der Standardmessfehler).

**sx**= Streuung des Merkmals in der Population (für Intelligenztests beträgt diese z.B. 15

-> hängt von  Skalierung (0 bis 10 vs. 0 bis 100) bzw. [Normierung](https://lms.uni-mainz.de/moodle/mod/resource/view.php?id=997403) der Testwerte ab (Intelligenzwerte werden tendenziell so normiert, dass der Mittelwert 100 und die Streuung 15 beträgt).

* KTT -> Standardmessfehler für alle Personen gleich
* Unabhängigkeitsaxiom: Messfehler unkorreliert (unabhängig von Eigenschaftsausprägung der Person)
* je größer Reliabilität, desto geringer Standardmessfehler
* Reliabilität = .80 -> 20% Messfehlervarianz

# Formel:

se = sx \* Wurzel ( 1-Rel.)

se = Standardmessfehler z.B. se = 4,74 => Testwerte würden im Durchschnitt 4,74 Punkte nach oben/unten schwanken

1. Rel. = **Fehlervarianz** = Anteil des Messfehlers an der beobachteten Varianz (Reliabilität = Anteil der Varianz der wahren Werte an der beobachteten Varianz)

* Wurzel (1 – Reliabilität) **= Fehlerstreuung**
* nur auf standardisierte Werte anwendbar
* Multiplikation mit sx (beobachtete Streuung über alle Personen) => Skalierung miteinbezogen
* kleine Messfehler wahrscheinlicher als große
* 1 SD Abweichung nach oben/unten => 68,27% der beobachteten Werte
* 2 SD Abweichung -> 95,45%
* 3 SD Abweichung -> 99,74%
* wahrer Wert nicht bekannt -> Standardmessfehler kann nur um beobachteten Wert gelegt werden
* Nutzung zur Definition eines Intervalls, innerhalb dem wahrer Wert mit bestimmter Wahrscheinlichkeit liegt z.B. 95%

= Konfidenzintervall/Vertrauensintervall

# Konfidenz-/Vertrauensintervall

* gibt den Bereich an, in dem der wahre Wert einer Person bei einer **vorher festgelegten Irrtumswahrscheinlichkeit α** liegt
* aufgrund der Messungenauigkeit der Verfahren keine Punktschätzung möglich
* Bestimmung eines Intervalls, innerhalb dessen der wahre Wert mit einer bestimmten Wahrscheinlichkeit vermutet wird
* je **höher Reliabilität**, desto **geringer der Standardmessfehler** und desto **schmaler das Konfidenzintervall**

# Formel:

CI = X+- zα/2 \*se

X = beobachteter Wert

zα/2 = 1,96 bei 95% Konfidenzintervall

* +-1,96 = z-Werte bei dem 95% der Fläche im Intervall liegt bzw. die jeweils 2,5% der Fläche abschneiden (zusammen 5%)
* Standardmessfehler ≠ 1 deshalb -> Multiplikation z.B. wenn se >1 Verteilung breiter
* wenn CI = 9,27 -> ausgehend vom beobachteten Wert 9,27 IQ-Punkte nach oben/unten

# Beispiel Intelligenz

* wahrer Wert =116
* entspricht überdurchschnittlicher Ausprägung ABER aufgrund von Messfehlern ist eine durchschnittliche Ausprägung nicht auszuschließen (Konfidenzintervall reicht bis in durchschnittlichen Bereich hinein)
* wenn Konfidenzintervalle zu weit keine präzisen Aussagen möglich

# Kritische Differenz

* Vergleich zwischen Personen
* Vergleich innerhalb einer Person

Ist die Differenz zwischen Testwerten hoch genug, damit von einem signifikanten Unterschied gesprochen werden kann?

DKrit = zα/2 \* seDiff

DKrit = Differenz die mindestens erreicht werden muss, um von einem signifikanten Unterschied zu sprechen

zα/2 -> 1,96 bei 95% Konfidenzintervall

* 10% Konfidenzintervall: 1,64
* 1% Konfidenzintervall: 2,57 (kritische Differenz ist größer als bei 5% Irrtumswahrscheinlichkeit)

seDiff = Standardmessfehler der Differenz

* Schwankungen der Differenzen abhängig von beiden Standardmessfehlern
* Schwankungen noch größer als die der einzelnen Werte

seDiff= Wurzel (se1^2 + se2^2)

# Entscheidungsregel

beobachtete Differenz ≤ kritische Differenz

* beobachtete Differenz kann auf Messfehler zurückgeführt werden => diagnostisch nicht bedeutsamer Unterschied

beobachtete Differenz > kritische Differenz

* beobachtete Differenz kann nicht durch Messfehler erklärt werden und ist überzufällig => diagnostisch bedeutsamer Unterschied

Bei identischer Reliabilität sowie Streuung des Merkmals (Vergleich zweier Personen auf derselbe Skala) können auch folgende Formeln verwendet werden:

* nicht bei Vergleich innerhalb einer Person auf zwei unterschiedlichen Skalen (unterschiedliche Reliabilitäten, Streuungen)

# Ein Bild, das Text enthält. Automatisch generierte Beschreibung

# Validität

= Übereinstimmung der Testergebnisse mit dem was der Test vorgibt zu messen

= Angemessenheit der Schlussfolgerung vom Testergebnis auf Verhalten außerhalb der Testsituation bzw. die Ausprägung eines bestimmten Merkmals

= wichtigstes Gütekriterium

* abhängig von Reliabilität und Objektivität -> notwendige aber keine hinreichenden Bedingungen
* durch beobachtbares Verhalten (Antworten auf einem Fragebogen etc. -> kann Gedanken, Gefühle und Verhaltensweisen die durch Konstrukt beeinflusst werden umfassen) auf latente Variable schlussfolgern
* Güte dieses Schlusses durch die Validität quantifiziert

# Validitätsarten

* Validität von Messungen -> lediglich Beschreibung von Personen
* Inhaltsvalidität, Konstruktvalidität
* Validität von Entscheidungen z.B. Studienbewerber, Personalauswahl
* Kriteriumsvalidität -> prädiktiv (liegt das zu erfassende Kriterium noch in der Zukunft (z.B. Tests zur Berufseignung und tatsächlicher Berufserfolg)) vs. Übereinstimmungsvalidität ( Prädiktor und Kriterium zeitgleich erhoben (z.B. Fragebogen zur Erfassung von Depression und Gespräch mit Diagnostiker)

# Inhaltsvalidität

= gegeben wenn der Inhalt des Tests bzw. der Items repräsentativ für das zu messende Merkmal ist

* Inhalt = Gesamtheit des Stimulusmaterials und aller Antwortalternativen
* Bestimmung der Repräsentativität idealerweise bereits während der Testkonstruktion
* Präzise Beschreibung des untersuchten Inhaltsbereichs (content domain) betrifft bereits Konstruktdefinition (präzise, klar, evidenzbasiert)
* Bestimmung des Teils des Inhaltsbereich der durch jedes einzelne Item gemessen wird (keine Items die sich keinem Inhaltsbereich zuordnen + kein Teil des Inhaltsbereichs der nicht durch Items abgedeckt wird)
* Vergleich der Struktur des Tests mit der des untersuchten Inhaltsbereichs
* evidence centered assessment design -> idealerweise bereits während Testkonstruktion systematische Argumentation für den Schluss der Itembeantwortung auf das untersuchte Konstrukt

Fragen:

* Ist der gesamte Inhaltsbereich durch Items erfasst?
* Lassen sich alle Items dem Inhaltsbereich zuordnen (oder sind irrelevante Items vorhanden)?
* Sind die Inhalte richtig gewichtet?
* Lassen sich Evidenzen für den Schluss vom Item auf das Konstrukt finden?

# Idealfall = Itemuniversum -> Gesamtheit aller Items zur Erfassung eines Merkmals aus denen repräsentative Auswahl für den Test getroffen wird

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

* Bestimmung der Inhaltsvalidität erfolgt nicht auf Basis empirischer Untersuchungen/nicht empirisch quantifiziert
* durch theoretische Argumente/Expertenmeinungen gestützt -> theoretisch/argumentativ, teilweise auch subjektiv geprägt

Konstruktvalidität = empirischer Beleg dafür das der Test das misst, was er zu messen vorgibt

# Cronbach, Meehl: Idealvorstellung eines nomologischen Netzes (umfasst Gesamtheit aller psychologischen Konstrukte)

* bisher kein vollständiges ausgearbeitetes nomologisches Netzwerk (formale Theorie)

= eher hoffnungslose Vision

# Vorgehen/Logik

1. Verhaltensindikatoren für Zielkonstrukt identifizieren
2. Identifikation weiterer Konstrukte und Beziehung zu Zielkonstrukt
3. Identifikation Verhaltensidikatoren für die weiteren Konstrukte und Beziehung zum Zielkonstrukt

* je mehr Überlappungen auf Ebene des beobachtbaren Verhaltens bzw. zwischen Verhaltensindikatoren desto größer Korrelation/Zusammenhang zwischen den Konstrukten
* Unabhängigkeit der Konstrukte wenn es keine Überlappungen gibt -> Fragebögen sollten nicht miteinander Korrelieren
* höchste Korrelationen mit den Konstrukten mit denen es die meisten Überlappungen gibt

# starker Ansatz -> ursprüngliches Ideal der Konstruktvalidität umgesetzt (formale Theorie)

# schwacher Ansatz (in Praxis häufiger) -> Validierung ohne formale Theorie aber gewisse Vorüberlegungen zu welchen Konstrukten Zusammenhänge bestehen könnten (vager, weniger präzise als starker Ansatz

# blinder Empirismus -> wahllose Korrelation mit anderen Testverfahren

* Korrelationsmuster im Nachhinein erklärt
* exploratives Erklären, wie Ergebnisse zu interpretieren sind
* liefert auch Information aber qualitativ geringer da exploratives Vorgehen

# Empirische Bestimmung der Konstruktvalidität

* häufigste Bestimmung durch:

1. Konvergente Validität liegt vor bei hohen Korrelationen des Tests mit anderen Verfahren die dasselbe oder ähnliche Konstrukte erfassen (viele Überlappungen bei Verhaltensindikatoren) *-> auch bei hohen negativen Korrelationen*
2. Diskriminante Validität liegt vor, wenn keine oder nur geringe Zusammenhänge zu Konstrukten bestehen, zu denen theoretisch kein Zusammenhang angenommen wird

# weitere Methoden (ergänzen sich)

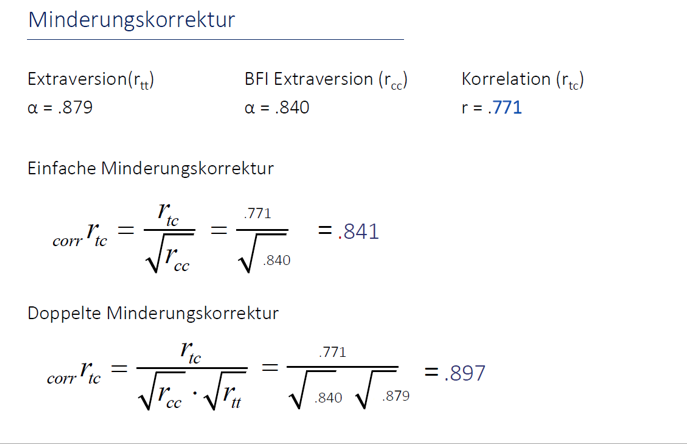
* **Korrelationen** -> Tests zum gleichen Konstrukt sollten positiv korrelieren
* **Gruppenunterschiede** -> Mittelwertsunterschiede zwischen bestimmten Gruppen, falls Test Konstrukt erfasst
* **Interne Struktur durch Faktorenanalyse** -> z.B. faktorenanalytischer Bestätigung der theoretisch angenommenen eindimensionalen Struktur des Konstrukts => faktorielle Validität
* **Veränderungen über die Zeit** -> durch Retest-Reliabilität quantifizierbar -> bestimmte Konstrukte die stabil sind (Persönlichkeit) oder sich verändern (Stimmung)
* **Untersuchung des Antwortprozesses** -> z.B. bei Rechentest Gefahr der Konfundierung mathematische Fähigkeiten mit Instruktionsverständnis -> Lösung lautes Denken
* **Experimentelle Interventionen** -> z.B. Veränderung Depressionswert nach Therapie

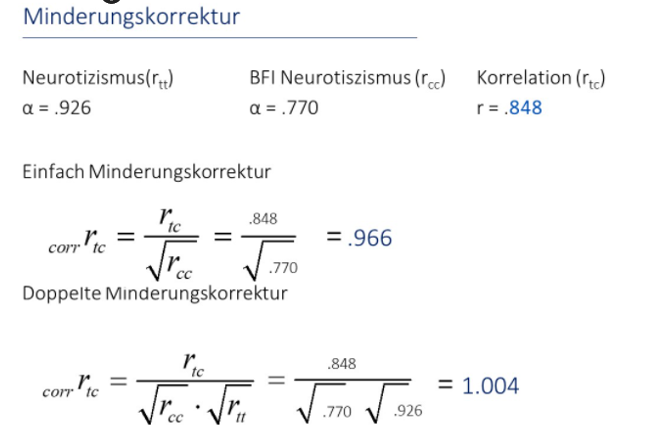
# Minderungskorrektur => KTT: Messfehler unkorreliert -> führt dazu dass:

* Korrelationen zwischen Tests fallen geringer aus als theoretisch angenommene Überlappungen vermuten lassen
* mangelnde Reliabilität reduziert Validität
* macht für eigenen Test Sinn: geringe Reliabilität legt nahe dass Validität des Schlusses vom Zielverfahren auf das Konstrukt begrenzt ist
* Ursache: Konstrukte werden durch andere Tests nicht messfehlerfrei gemessen -> reduzieren Korrelation und damit die Validität des Zielfragebogen unverschuldet
* Messfehler des Validierungsverfahrens aus Validität herausrechnen (Annahme Reliabilität des Validierungsverfahrens = 1)
* Korrelation zwischen Zielfragebogen t und Validierungsverfahren c und geteilt durch Wurzel aus Retest-Reliabilität (rcc) des Validierungsverfahren (auch möglich durch Einsetzung von Cronbachs Alpha oder Testhalbierungreliabilität)
* doppelte Minderungskorrektur: Validität noch stärker nach oben korrigiert -> Annahme unter Bedingung dass beide Verfahren messfehlerfrei messen ( z.B. sinnvoll wenn kürzere Version des Zielfragebogens erstellt werden soll z.B. BFI-Gruppe -> Nachweis das kürzere Version und Originalversion dasselbe Konstrukt erfassen => Korrelation sollte r= 1 betragen) –
* Validität kann nach doppelter Minderungskorrektur auch Werte größer als 1 annehmen (sollte eigentlich nicht möglich sein) = Überkorrektur (Cronbachs Alpha unterschätzt Reliabilität tendenziell wenn Konstrukte sehr hetrogen sind -> geringe Reliabilitäten führen zu einer stärkeren Korrektur nach oben bei der Minderungskorrektur und damit zu einer Überkorrektur (kann auch mit anderen Reliabilitätsmaßen passieren da Reliabilität nur geschätzt wird -> Korrektur führt immer auch Verzerrungen in die eine oder andere Richtung DESHALB ist auch immer die unkorrigierte Validität anzugeben)

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung





# Multitrait-Multimethod-Ansatz

**Konvergente** Validität liegt vor, wenn die Messungen desselben Konstrukts mit unterschiedlichen Methoden hoch miteinander korrelieren.

**Diskriminante** Validität liegt vor, wenn die Messung verschiedener Konstrukte mit derselben Methode ODER verschiedenen Methoden nicht oder nur gering korrelieren

Methodeneffekt

* wenn zwei Konstrukte mit derselben Methode erfasst werden, kann Teil der Korrelation eventuell auf die gemeinsame Methode zurückgeführt werden
* MTMM-Ansatz erlaubt eine Schätzung dieses Methoden-Bias

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

* 2 Konstrukte mit 3 Methoden erfasst => für jede Person 6 Testwerte
* jedes Konstrukt mit allen 3 Methoden gemessen
* Hauptdiagonale -> gleiches Konstrukt mit gleicher Methode erfasst (Monotrait-Monomethod) => müsste eigentlich r =1 sein ABER geringerer Wert durch Korrelation von Retest-Reliabilitäten
* konvergente Validität -> gleiches Konstrukt; unterschiedliche Methoden (Monotrait-Multimethod)
* diskriminante Validität -> unterschiedliche Konstrukte mit gleicher Methode (Multitrait-Monomethod) oder unterschiedlichen Methoden (Multitrait-Multimethod)

# Nachweis der konvergenten Validität

* Monotrait-Multimethod-Koeffizienten sollten statistisch signifikant sein
* aber auch abhängig von SP-Umfang (bei großen Stichproben auch kleinere Zusammenhänge signifikant
* Festlegung einer Mindestkorrelation in Abhängigkeit der Reliabilitäten
* Orientierung an Literatur (Validierungsstudien, Meta-Analysen)
* *Wie hoch korrelieren gleiche Konstrukte, die mit unterschiedlichen Methoden gemessen wurden im Schnitt?*
* bei Vergleich von Selbst- und Fremdbericht keine Korrelation von r= 1 erwartbar, da neben Validität auch von Unterschieden in der Selbst- und Fremdwahrnehmung abhängig

# Nachweis der diskriminanten Validität

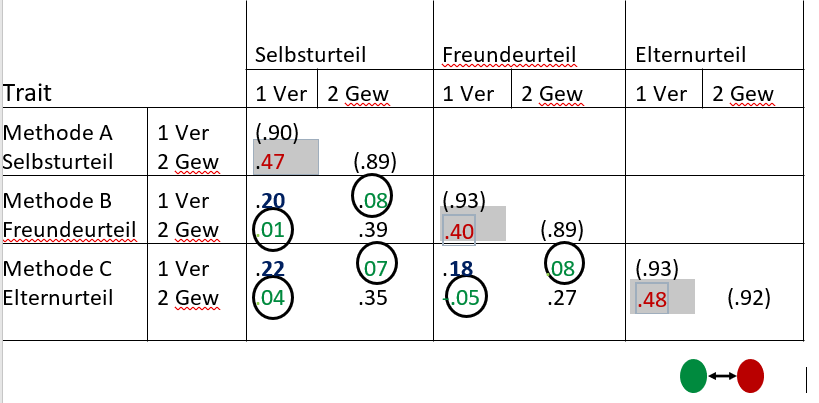
* Multitrait-Multimethod-Koeffizienten **sind bei allen Vergleichen niedriger**, als konvergente Validitäten
* Multitrait-Monomethod Koeffizienten sollten niedriger sein als konvergente Validitäten (Monotrait-Multimethod)
* wenn Multitrait-Monomethod Koeffizienten größer als konvergente Validitäten liegt starker Methodeneffekt vor oder diskriminante Validität ist tatsächlich eingeschränkt
* Muster der Merkmalskorrelationen sollten zwischen sowie innerhalb der Methoden konstant sein -> z.B. Monotrait-Multimethod- Koeffizienten immer größer als Multitrait-Monomethod Koeffizienten

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

# Nachweis eines Methodeneffekts

* Methodeneffekt liegt vor, wenn Multitrait-Monomethod Koeffizienten statistisch signifikant höher als Multitrait-Multimethod-Koeffizienten
* Methodeneffekt z.B. durch Antworttendenzen
* wenn Methodeneffekt festgestellt können Multitrait-Monomethod Koeffizienten nicht problemlos zur Beurteilung der diskriminanten Validität herangezogen werden



# Pro

+ strukturierter Ansatz zur Beurteilung der Validität des Verfahrens

+ Methodeneffekte können festgestellt werden

# Kritik

* Beurteilung der Validität auf Basis der Häufigkeitszählung vieler Einzelvergleiche vorgenommen
* Reliabilitäten unberücksichtigt
* keine exakten Entscheidungsregeln -> individuelle Entscheidungen bei Beurteilung der Validität

DESHALB

* Prüfung des MTMM-Designs durch konfirmatorische Faktorenanalyse durch Strukturgleichungsmodelle = Analyse der gesamten Korrelationsstruktur auf einmal
* ermöglicht aggregierte Ergebnisse
* Trennung des Traits-, Methoden- und Messfehleranteils möglich
* Überprüfung der Eindimensionalität einzelner Konstrukte (durch MTMM-Ansatz nicht möglich)

# Objektivität

* Objektivität = Ausmaß, in dem die Ergebnisse eines Tests unabhängig von der Person des Untersuchungsleiters sind
* gesichert durch Standardisierung der einzelnen Phasen des diagnostischen Prozesses (gleiche Bedingungen damit Durchführung nicht selbst Varianz produziert)

Arten von Objektivität:

* Durchführungsobjektivität
* Auswertungsobjektivität
* Interpretationsobjektivität

# Durchführungsobjektivität

Durchführungsobjektivität

* Sicherung durch maximale Standardisierung der Testsituation (festgelegte Reihenfolge/Formulierung der Items, minimale Interaktion mit Testleiter, Fragebögen unter gleichen Bedingungen (Raum…)
* bei Onlineerfassung eventuell nicht alle Bedingungen gleich (problematisch bei individueller Leistungserfassung
* z.T. Interaktion mit Leiter erforderlich (klinische Einrichtungen) -> nicht online sondern in konstantem Rahmen -> Leiter muss hinreichend geschult sein -> WAIS-IV mit standardisierten Instruktionen (Einarbeitung -> Nachfragen…Anweisungen müssen wörtlich auswendig gelernt/dürfen nur wiederholt nicht erklärt werden)
* Beispiel für standardisierte Testinstruktion:

Instruktion für den Anwender: Aus dem HAWIE (Wechsler, 1964):

# Auswertungsobjektivität

* gegeben bei eindeutiger Quantifizierung des Verhaltens
* geringer bei freiem als gebundenem Antwortformat
* gebunden: mögliche Fehler durch Recodierung… -> Schablonen die auf Fragebogen aufgelegt werden können oder bestimmte Auswertungssoftwares/Auswertung über Syntax
* frei -> Hinweise z.B. welche Antworten noch/nicht mehr gültig ABER eventuell auch nicht aufgeführte Begriffe -> Listen müssen immer weiterentwickelt werden

Aus dem „Lern- und Gedächtnistest (LGT 3)“(Bäumler, 1974)

In der 3. Aufgabe werden dem Probanden 20 Gegenstände gezeigt (bildlich), z.B.

Kleiderbügel, Hammer, Roller.

*Zur Auswertung werden folgende Hilfen angeboten:*

*Richtig Noch gültig Ungültig*

*Kleiderbügel Aufhänger, Haken Bogen, Bumerang*

*Hammer Schlegel Werkzeug*

*Roller Zweirad Fahrrad, Rad*

* quantitative Bestimmung: Vergleich der Ergebnisse von mindestens zwei Auswertern
* Bestimmung der Auswertungsobjektivität mittels

**Intraklassenkorrelation** (Maß für Beurteilerübereinstimmung)

# Interpretationsobjektivität

Interpretationsobjektivität

– Gegeben, wenn aus **gleichen Scores** von **verschiedenen Auswertern identische Schlüsse** gezogen werden

- Einordnung (Merkmalsbeschreibung als durchschnittlich…) und festlegen von cut-off Werte (geeignet/nicht geeignet)

- Normwerte z.B. IQ

# Zusammenfassung Gütekriterien

Wechselbeziehung zwischen den Gütekriterien Objektivität, Reliabilität und Validität (nach Lienert)

Objektivität

Interpretation

Auswertung

Durchführung

Reliabilität

innere Konsistenz + Tes

Parallel

-

und

Retest

Validität

(

kriteriumsbezogen

)

Zulänglichkeit des

Kriteriums

Stabilität des

Persönlichkeits

-

merkmals

* Auswertungs/Durchführungsobjektivität, Stabilität (Retest)/Heterogenität (interne Konsistenz) des Merkmals kann sich negativ auf Reliabilität auswirken
* Alpha Maximierung kann Reliabilität erhöhen aber reduziert Validität (Verdünnungsparadox)
* Kriterien müssen hohe Validität aufweisen z.B. Studienerfolg (Noten, Dauer…)

# Beurteilung der Höhe von Testkennwerten (nach Weise, 1975)

|  |
| --- |
| **Kennwert Kürzel Niedrig Mittel Hoch** |
| Schwierigkeit P\*)2 <.20 .20 - .80 > .80  Trennschärfe (korrigiert) ritc < .30 .30 - .50 > .50  Objektivität (Auswerter) rk < .60 .70 - .90 > .90  Reliabilität rtt < .80 .80 - .90 > .90  Validität \*)1 (unkorrigiert) rtc < .40 .40 - .60 > .60  Eichstichprobe N < 150 150 - 300 > 300 |
| \*)1 Validitätskoeffizient besagt wenig über die Bedeutung des Tests, wenn man nur den absoluten Betrag bewertet. Beachtet werden muss auch der Beitrag, den ein Test zur Lösung einer gegebenen Fragestellung leisten kann (z.B. Basis-, Selektionsrate und Validität)  \*)2 p = relativer Anteil von Probanden, die ein Item „richtig“ beantworten. |

Beispiel für Richtlinien zur Bewertung der Reliabilität sowie des Umfangs der Normstichprobe (Evers, 2001)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Reliabilität1** | **Umfang Normstichproben** | |
|  | Niveau 1 | Niveau 2 Niveau3 | Niveau 1 Niveau2 | Niveau3 |
| unzureichend ausreichend gut | < .80  .80 - .90  > .90 | < .70 < .60  .70 - .80 .60 - .70  > .80 > .70 | < 300 < 200  300 – 400 200 – 300  > 400 > 300 | < 100  100 – 200  > 200 |

Anmerkung: Niveaus (1) Tests für wichtige Entscheidungen auf der individuellen Ebene (z.B. Personalauswahlentscheidungen), (2) Tests für weniger bedeutsame Entscheidungen auf individueller Ebene (z.B. Fortschrittskontrolle), (3) Tests für Untersuchung auf Gruppenniveau. 1 Für Paralleltest-Reliabilität, interne Konsistenz, Test-Retest-Reliabilität und Interrater-Reliabilität

# Normierung

# Zwei Formen der Testwertinterpretationen

Normorientierte Testwertinterpretation

Kriteriumsorientierte Testwertinterpretation

# 1. Normorientierte Testwertinterpretation

Durch Normierung – auf Basis einer großen und repräsentativen Stichprobe – erhält man ein **Bezugssystem**, um individuelle Testwerte einordnen bzw. interpretieren zu können.= Testeichung

1. Definition der Zielpopulation
2. Wahl des Erhebungsdesigns

# Globale Repräsentativität

(bestenfalls über Panels aber teuer und deshalb in Praxis selten): wenn die Zusammensetzung der Stichprobe hinsichtlich **aller möglichen Faktoren (Geschlecht, Alter…)** mit der Populationszusammensetzung übereinstimmt (*Zufallsstichprobe aus der definierten Zielpopulation*)

* zunächst in verschiedenen Personengruppen z.B. Studenten Daten erhoben -> später Erweiterung auf andere Gruppen => zusammengesetzte Stichproben (Patchwork-SP) -> nicht unbedingt repräsentativ bzgl. Zielpopulation
* SP hinterher so zusammengestellt dass spezifische Repräsentativität bzgl. einiger wichtiger Merkmale der Zielpopulation hergestellt z.B. Alter Geschlecht…

# Spezifische Repräsentativität:

wenn die Zusammensetzung der Stichprobe hinsichtlich derjenigen Faktoren der Populationszusammensetzung repräsentativ ist, die mit dem Testmerkmal in irgendeiner Weise zusammenhängen (z.B. Geschlecht -> Persönlichkeit, Alter, Bildungsgrad)

Der individuelle Testwert wird in einen **Normwert transformiert**, der es ermöglicht, den Testwert in Bezug zu einer **Referenzgruppe** zu interpretieren.

Warum?

*Rohwerte (z.B. Summenscores) sind wenig aussagekräftig!*

* Test 1: 80 von 100 Punkten; Test 2: 60 von 100 Punkten
* Der zweite Test kann für eine höhere Leistung/Merkmalsausprägung (relativ zur Referenzgruppe)stehen als der erste Test ABER auch durch unterschiedliche Schwierigkeiten der beiden Tests zustandekommen (keine Aussage über MW in der Referenzpopulation)

Variabilitäts – oder Abweichungsnormen

Prozentrangnormen

Äquivalenznormen

# Variabilitäts- und Abweichungsnormen

* Abweichung vom Mittelwert der Vergleichsstichprobe in Einheiten der Standardabweichung
* Voraussetzung – Normalverteilung

1.1 Variabilitäts- und Abweichungsnormen

Rohwertbildung z.B. Summenscore, Mittelwert

|  |  |
| --- | --- |
| z-Standardisierung | Umrechnung in einen z-Wert |

(Transformation) z-Wert \* Multiplikator + Konstante

Normwert IQ-Skala; Z-Wert (Standardwert);

T- Wert; C-Wert; Stanine-Werte

(z-Standardisierung)

* z-Werte mit MW = 0 SD = 1

# Rohwerte sind abhängig von der Anzahl der Aufgaben (bei Summierung) sowie der Skalierung (1 – 7; 0 – 100 usw.)  schwer zu interpretieren

1, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 9, 10

(M = 5.35; SD = 2.31)

Variabilitäts- und Abweichungsnormen (z-Standardisierung)

Rohwerte (X):

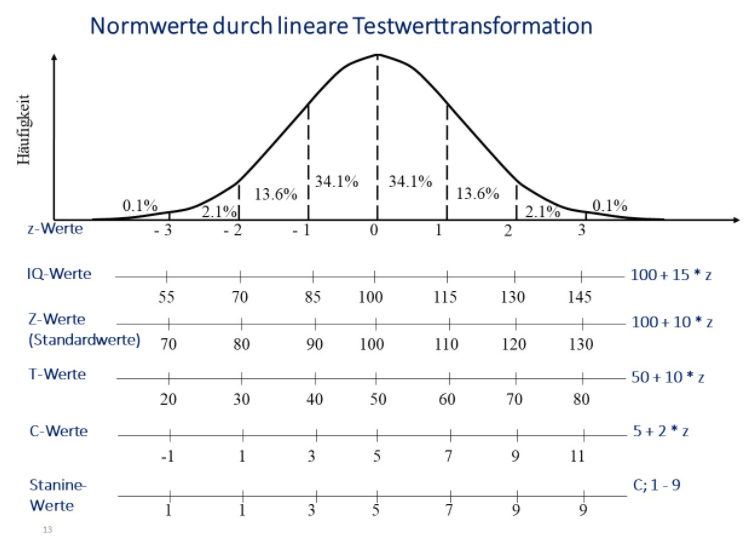
1, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 9, 10

(M = 5.35; SD = 2.31)  (M = 0.00; SD = 1.00)

Rohwerteverteilung Standardnormalverteilung

(X – M) / sX

|  |
| --- |
| 1.1 Variabilitäts- und Abweichungsnormen (z-Standardisierung)  Rohwerte (X):  1, 2, 2, 3, 3, 4, 4, 4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8, 8, 9, 10  -1.88, -1.45, -1.02, -0.58, -0.15, 0.28, 0.71, 1.15, 1.58, 2.01 -1.45, -1.02, -0.58, -0.15, 0.28, 0.71, 1.15 -0.58, -0.15, 0.28, 0.71  -0.15, 0.28 |
| z-Wert = (Rohwert – MW )/SD  z Wert 0 = Durchschnitt (im Beispiel Rohwert 5)  z Wert 1 = 1 SD über MW |
|  |



# Normwerte Verteilung

Stanine Werte => differenzieren weniger stark in Extrembereiche im Vergleich zu C-Werten -> wenn Verfahren gute Differenzierung in diesen Bereichen nicht zulässt

T-Werte können innerhalb einer Standardabweichung besser differenzieren als C-Werte => Test mit sehr guter Reliabilität (genauere Abstufungen -> weniger Personen die auf einen Zahlenwert fallen) => Vorzug von Normwerten ggü. Standardnormalverteilung (feiner abgestuft als z-Werte)

IQ-Werte

Z-Werte Intelligenztests

T-Werte Persönlichkeitstests

C-Werte gröbere Norm, wenn feine Unterteilung

Stanine-Werte unangemessen( z.B. wegen mangelnder

Reliabilität

1.1 Variabilitäts- und Abweichungsnormen

# Beispiel NEO-PI-R: 5 Altersgruppen, Geschlecht, Gesamtstichprobe

(\*repräsentativ hinsichtlich Alter, Geschlecht, Bildungsstand)

-> Gesamtstichprobe größer bevölkerungsrepräsentative SP

Stichprobe nicht repräsentativ da nichtrepräsentative SP erhoben wurden z.B. studentische SP => typische Patchwork-SP

* repräsentative Einordnung ermöglicht durch bevölkerungsrepräsentative Zusammensetzung der Gesamtstichprobe ABER dadurch reduziert sich Referenzgruppe auf N= 871 (Personen entfallen z.B. durch zu viele junge Leute in Basisdaten…im Vergleich zu Gesamtbevölkerung)

|  |  |  |  |
| --- | --- | --- | --- |
| **NEO-PI-R** | **Beschriftung der Kopfzeile des Profils** | **N** | **Farbe der**  **Profilkopfzeile** |
| Form S 1.  2.  3.  7. | 1. Gesamtstichprobe 2. bevölkerungsrepräsentative Gesamtstichprobe\* 3. Männer /16 – 20 Jahre 4. Frauen / 16 – 20 Jahre 5. Männer / 21 – 24 Jahre 6. Frauen / 21 – 24 Jahre   …   1. bevölk.repr. Stichprobe Männer 2. Bevölk.repr. Stichrpobe Frauen | 11 724  871  480  1686  1358 1925  423  448 | blau Grün  blau (heller als 1a) blau (wie 2a)  blau (heller als 2) blau (wie 3a)  grün (heller als 1b) grün (wie 7a) |

Normierung

**1.1**

**Variabilitäts**

**-**

**und Abweichungsnormen**

Beispiel NEO

-

PI

-

R: 5 Altersgruppen, Geschlecht, Gesamtstichprobe

(

\*repräsentativ hinsichtlich Alter, Geschlecht, Bildungsstand

)

-

Unterschiedliche Profilbögen für unterschiedliche Stichproben

Gesamtstichprobe; Männer; Frauen; Männer

(

16

-

20

; Frauen

16

-

20

usw.)

-

Angabe der

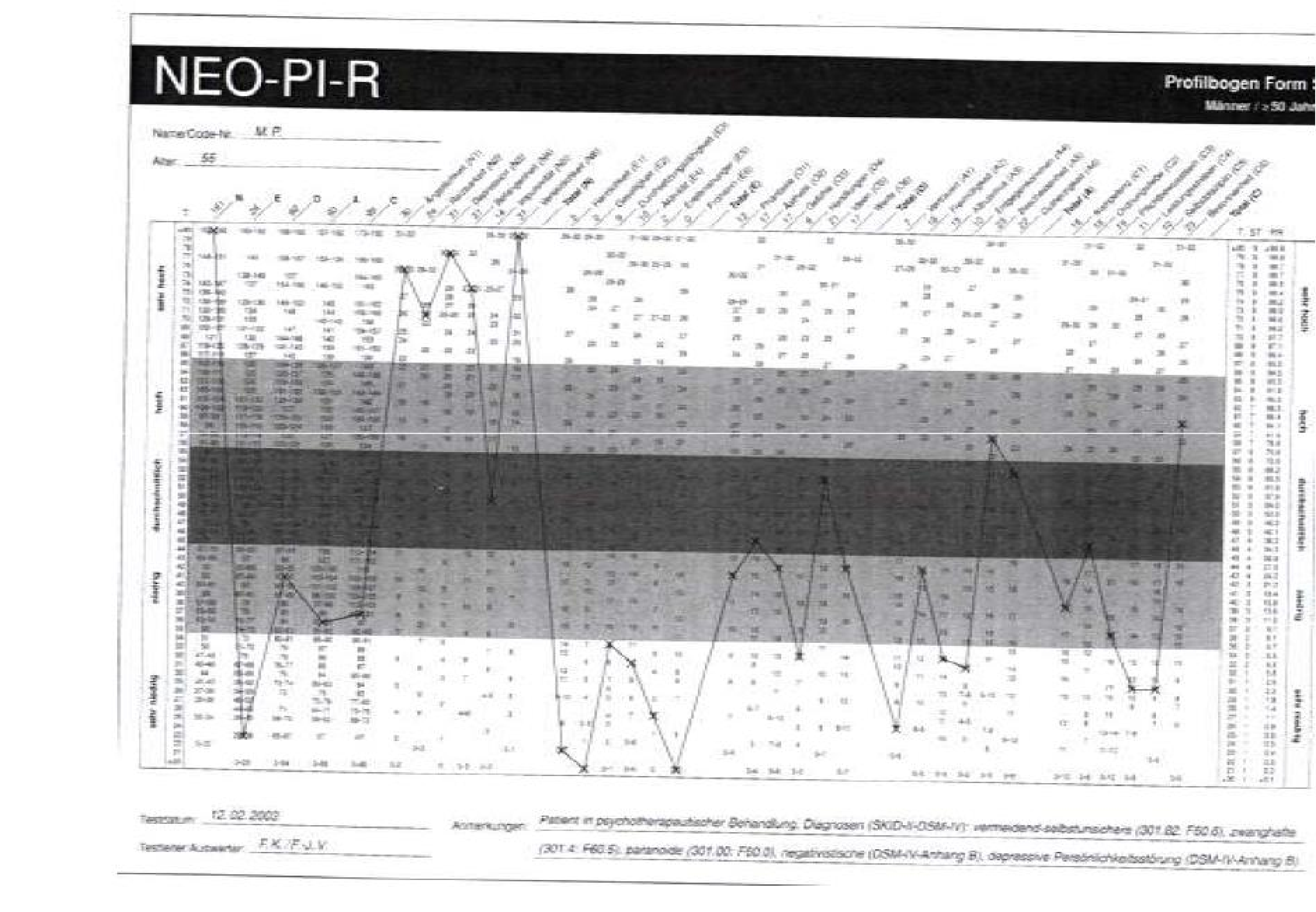
Rohwert

(

Rohwerterange) und entsprechender T

-

Werte, Prozentränge



# Was tun, wenn im Test keine Normtabellen angegeben?

– Wenn Mittelwert und Standardabweichung im Manual angegeben sind, dann kann der Rohwert über die z-Standardisierung in jeden beliebigen Normwert transformiert werden

Rohwert: 35 (M = 24.00; SD = 8.00) z-Wert : 1.375

T-Wert: 50 + 10\*1.375 = 63.75 = 64

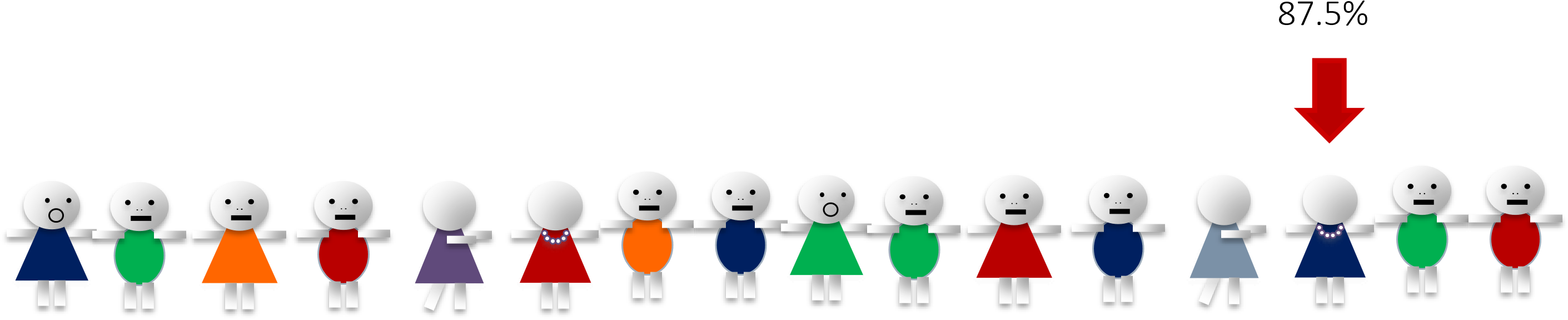
1.1 Variabilitäts- und Abweichungsnormen

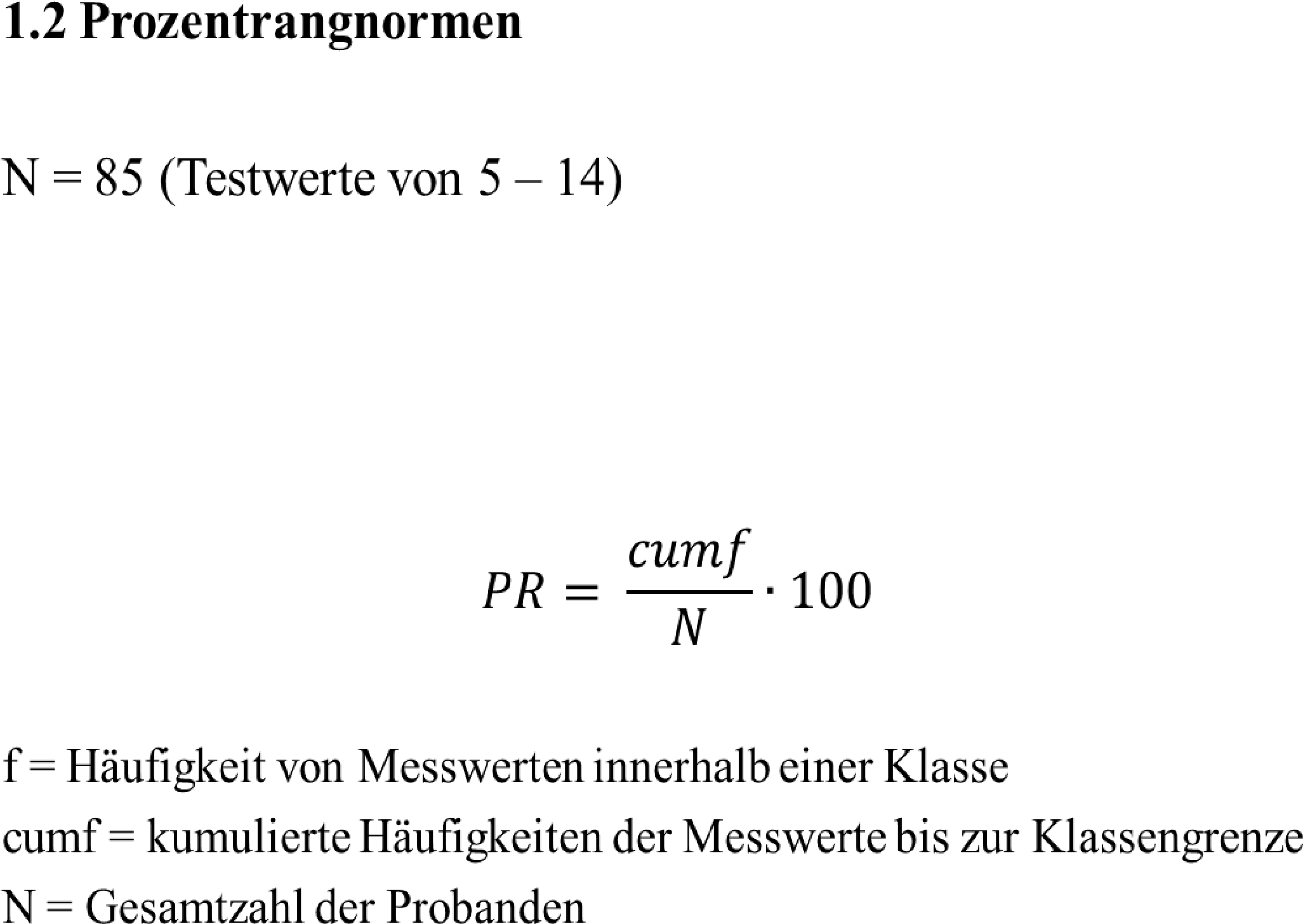
Wenn ein Merkmal (empirisch) nicht normalverteilt ist, verbietet sich die Transformation in Normwerte (es sei denn, die Annahme für Normalverteilung ist begründbar.

* Rohwerte über Flächentransformation in Normalverteilung überführen
* anschließend Transformation in Normwerte

## Prozentrangnormen

* Rohwerten wird die relative Position in einer Bezugsgruppen zugeordnet
* Anteil der Normstichprobe, der gleich große oder kleinere Werte erreicht hat (in Prozent)
* auch bei nicht normalverteilten Rohwerten möglich





•

Testwert

5

6

7

8

9

10

11

12

13

14

f

1

0

6

10

19

24

14

6

4

1

cumf

1

1

7

17

36

60

74

80

84

85

1.2 Prozentrangnormen

N = 85 (Testwerte von 5 – 14)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Testwert | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| F | 1 | 0 | 6 | 10 | 19 | 24 | 14 | 6 | 4 | 1 |
| cumf | 1 | 1 | 7 | 17 | 36 | 60 | 74 | 80 | 84 | 85 |

PR = 17/85 \* 100 = 20 % (Prozentrang für Testwert 8)

f = Häufigkeit von Messwerten innerhalb einer Klasse

cumf = kumulierte Häufigkeiten der Messwerte bis zur Klassengrenze

N = Gesamtzahl der Probanden

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| z-Werte  Prozentrang | - 3  0.1% | - 2  2.3% | - 1  15.9% | 0 50% | 1  84.1% | 2  97.7% | 3  99.9% |
| T-Werte | 20 | 30 | 40 | 50 | 60 | 70 | 80 |
| IQ-Werte | 55 | 70 | 85 | 100 | 115 | 130 | 145 |

1.2 Prozentrangnormen (direkter Zusammenhang mit Normwerten bei gegebener Normalverteilung)

Häufigkeit

1.2

Prozentrangnormen

1

2

3

-

1

-

2

-

3

Häufigkeit

0

50

%

%

15.9

z

-

Werte

%

2.3

%

0.1

84.1

%

97.7

%

%

99.9

Prozentrang

PR

-

Diff = 2.2 %

Diff=1

Diff=1

PR

-

Diff = 34.1 %

Problem: Interpretierbarkeit – Normwerte sind intervallskaliert, Prozentränge hingegen nicht. Daher sind Prozentrangdifferenzen nicht wie Normwertdifferenzen zu interpretieren.

Prozentrangdifferenzen zwischen z-Werten 0 und 1 (34,1%) wesentlich größer als zwischen 2 und 3 (2,2%) Abweichung aber gleiche Veränderung in der Merkmalausprägung/Leistung

im mittleren Bereich führen bereits kleine Verbesserungen zu großen Prozentrangsprüngen (weniger Personen mit Werten in Extrembereichen -> Glockenkurve)

1.3 Äquivalenznormen

* Zuordnung der Rohwerte zu bestimmten Referenzgruppen z.B. Alter oder Reifeabschnitt, für die die betreffende Leistung besonders typisch ist.
* älteste Art der Normierung
* heutiger Einsatz: Bestimmung des Altersniveau bei z.B. Personen mit geistiger Beeinträchtigung

## Äquivalenznormen

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Alter** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Anzahl Aufgaben | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 12 |
| Anzahl gelöst | 12 | 12 | 12 | 12 | 8 | 5 | 3 | 0 |

Proband: Alter = 7 Jahre 3 Monate (87 Monate)

Intelligenzalter = 6 Jahre + 8 Monate + 5 Monate + 3 Monate

= 7 Jahre 4 Monate (88 Monate)

IQ = Intelligenzalter / Alter = 88/87 = 1.01 \* 100 = 101

# Kriteriumsorientierte Testwertinterpretation

Interpretation in Bezug auf ein spezifisches inhaltliches Kriterium (z.B. diagnostische Aussage – Depression **ja/nein) => es soll Entscheidung getroffen werden**

* Definition eines Schwellenwerts
* Bezug des Testwertes auf ein externes Kriterium z.B. durch umfangreiches klinisches Interview = Goldstandard
* ROC-Analyse zur Bestimmung des Schwellenwertes

|  |  |  |
| --- | --- | --- |
|  |  | **Klassifikation/Entscheidung** |
|  |  | - + |
| **Kriterium** | + | Falsch Negativ Richtig Positiv  (FN) (RP) |
|  | - | Richtig Negativ Falsch Positiv  (RN) (FP) |

RP = valide positiv/ richtig positiv

RN = valide negativ/richtig negativ

**Abgelehnt**

**Akzeptiert**

**e**

**Eignung**

**Geeignet**

**Ungeeignet**

2.1

Bezug des Testwertes auf ein externes Kriterium

–

ROC

-

Analyse zur Bestimmung des Schwellenwertes

RP = valide positiv/ richtig positiv

RN = valide negativ/richtig negativ

**Klassifikation/Entscheidung**

**nein**

**ja**

**Tatsächlich**

**depressiv**

**(**

**klinisches**

**Interview)**

**ja**

Falsch Negativ

(

)

FN

Richtig Positiv

(

RP

)

**nein**

Richtig Negativ

(

)

RN

Falsch Positiv

(

FP

)

Schwellenwert zu niedrig -> viele falsch Positive; weniger falsch Negative durch Selbstberichtstest

Schwellenwert zu hoch -> viele falsch Negative; weniger falsch Positive

55

2.1 Bezug des Testwertes auf ein externes Kriterium

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

Sensitivität erhöht -> Schwellenwert niedriger -> Spezifität niedriger

ROC-Analyse  Schwellenwert, der ein optimales Gleichgewicht zwischen Sensitivität und Spezifität herstellt.

(Fehler in beide Richtungen in der Summe minimieren)

Bezug des Testwertes auf ein externes Kriterium

# ROC-Analyse  Schwellenwert, der ein optimales Gleichgewicht zwischen Sensitivität und Spezifität herstellt.

Beispiel: Entwicklung Depressionsfragebogen

1. Klassifikation: klinisches Interview (Depressive vs. nicht Depressive) = Kriterium das Probanden zuverlässig in depressiv vsm. nicht-depressiv aufteilt
2. Erhebung der Testwerten
3. Berechnung der Sensitivität und Spezifität für jeden potentiellen Schwellenwert
4. Optimaler Schwellenwert bei der Sensitivität + Spezifität maximal sind

ROC-Kurve:

Spezifität = hohe Werte niedrige Spezifität; niedrige Werte hohe Spezifität

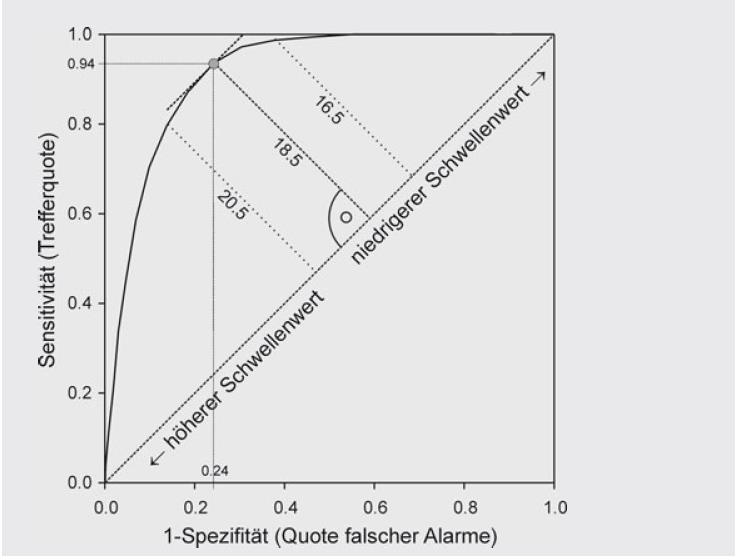
Verschiebung nach oben -> Senkung Schwellenwert -> Zunahme Sensitivität

Verschiebung nach unten -> Erhöhung Schwellenwert -> Zunahme Spezifität

Punkt an dem Optimum erreicht = Scheitelpunkt der Kurve

Beispiel: Schwellenwert 18,5 Sensitivität 0,94 (94% der tatsächlich depressiven auch tatsächlich als depressiv klassifiziert) Spezifität 0,76 (76% der nicht Depressiven als nicht depressiv klassifiziert; 24% fälschlicherweise als depressiv klassifiziert)

* in der Praxis Fälle wo hohe Sensitivität wichtiger als hohe Spezifität und umgekehrt z.B. tatsächlich Infizierte entdecken -> Schwellenwert herabsetzen -> Sensitivität erhöht



# Bezug des Testwertes auf Aufgabeninhalte

Testwert im Sinne des Ausmaßes der Erfüllung eines z.B. Lernziels zu interpretieren.

– Anteil gelöster Aufgaben => Anwendung bei Lernziel- oder Leistungstests