LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Vorlesung Testtheorien

Dr. Tobias Constantin Haupt, MBA

Sommersemester 2007

Objektivität: Sie muss von verschiedenen Auswertern gleichermaßen als „richtig“ oder „falsch“ bewertet werden

Reliabilität: Gleiche Antworten bei kurzeitiger Testwiederholung muss erwartet werden

Validität: Eine Aufgabe ist dann valide, wenn sie von Pbn mit starker Merkmalsaufprägung häufiger richtig beantwortet wird als von Pbn mit geringer Merkmalsausprägung (vgl. Trennschärfe)

Inhaltliche Kriterien einer guten Testaufgabe

Die Aufgabe als inhaltliche Ganzheit soll einen wesentlichen Aspekt des untersuchten Merkmals betreffen

Jede Aufgabe soll von jeder anderen inhaltlich unabhängig sein. Die Lösung einer Aufgabe darf keinen Hinweis auf die Lösung einer anderen Aufgabe enthalten und schon gar nicht von deren Lösung abhängig sein

Jede Aufgabe soll, soweit dies mit dem Testziel vereinbar ist, speziell, konkret und wirklichkeitsnah gestaltet sein und nicht allgemein, abstrakt und wirklichkeitsfern

Regeln für den sprachlichen Aufbau einer Testaufgabe

- Vermeide mehrdeutige Begriffe (u.a. „oft“)
- Vermeide Begriffe, die nur einem Teil der Pbn geläufig sind
- Jede Aufgabe soll nur einen Aspekt oder Gedanken enthalten, d.h. keine durch „und“ verbundenen Aussagen
- Möglichst positive Aussagen, Fragen oder Formulierungen, Vorsicht: doppelte Verneinung)
- Vermeide Verallgemeinerungen jeder Art, besonders solche, die nur zeitweise Gültigkeit haben
- Vermeide umständliche Länge oder telegrafische Kürze im Aufgabentext
- Prüfe die Eindeutigkeit der Frage an einer kleinen Teststichprobe („Fragentext ist eindeutig / nicht ganz / unklar“)

- Eigentlich nur problematisch bei Persönlichkeitstests, die keine im engeren Sinne richtigen oder falschen Antworten enthalten können

- Bei gestuften Antwortskalen der Form „immer – manchmal – selten – nie“ wird die im Sinne des Testgegenstandes beste Antwort am stärksten gewertet.

- Items sind die *kleinsten Elemente eines Tests*, von denen damit letztlich seine Qualität abhängt.

- Im Verlauf der Itemanalyse werden die psychometrischen Itemeigenschaften als Kennwerte bestimmt und anhand vorgegebener Qualitätsstandards beurteilt.

<u>Im Einzelnen</u> umfaßt eine Itemanalyse meist

- die Analyse der Rohwertverteilung (diese sollte in der Regel aufgrund ihrer inferenzstatistisch wünschenswerten Eigenschaften normalverteilt sein),

und die Berechnung von

- Itemschwierigkeit,
- Trennschärfe,
- Homogenität, sowie einer
- Dimensionalitätsüberprüfung.

- *Ziel* ist es in der Regel, diejenigen Items mit den besten psychometrischen Eigenschaften auszuwählen.
- Die folgenden Itemkennwerte sind sowohl psychometrische Charakteristika von Items als auch Gütekriterien zur Auswahl von für einen Test besonders geeigneten Items.

Itemschwierigkeit I

Definition: Die **Schwierigkeit** eines (einzelnen) Items (in einer gegebenen Stichprobe!) gibt an, wie groß der relative (prozentuale) Anteil von Probanden ist, die ein Item im Sinne höherer Merkmalsausprägungen beantworten.

Je mehr Versuchspersonen einer SP das Item in Merkmalsrichtung beantworten, desto geringer die Schwierigkeit.

Itemschwierigkeit II

- Begriff, der aus den Leistungstests stammt und für Rating - Skalen übernommen wurde
- Wenn eine Aufgabe von vielen Personen gelöst wird, gilt sie als leicht. Viele Personen haben dann eine hohe Punktzahl in der Aufgabe erzielt, der Mittelwert der Punkte in der Aufgabe über alle Personen ist hoch. Dieser Mittelwert wird auch Schwierigkeit genannt.
- Man sagt, daß die Aufgabe bei hohem Mittelwert eine geringe Schwierigkeit hat, bei geringem Mittelwert eine hohe Schwierigkeit.

Itemschwierigkeit und Differenzierungsfähigkeit

Mittlere Schwierigkeitskoeffizienten (um .50):

- Bedeutet größtmögliche Streuung der Itembeantwortungen über die Versuchspersonen und damit auch größtmögliche Differenzierung über die GesamtSP hinweg.
- Große Merkmalsstreuungen begünstigen (im Sinne einer notwendigen Bedingung) hohe Korrelationen, was wiederum eine günstige Voraussetzung für Trennschärfe des Items und Homogenität der Skala ist.

Extreme Schwierigkeitskoeffizienten (.05-.10, .90-.95):

Hätte man nur Items mit mittlerer Schwierigkeit, so würde der Test nur zwischen den zwei Gruppen Löser versus Nichtlöser differenzieren.

Um auch zwischen Versuchspersonen mit extremeren Merkmalsausprägungen differenzieren zu können, bedarf es zusätzlich Items mit extremeren Schwierigkeitskoeffizienten. Gleichwohl führt dies zu Einbußen an Trennschärfe und Homogenität.

Itemselektion – Wonach geht man?

- Inhaltliche Kriterien
- Statistische Kriterien
- Sonstige Kriterien

Itemselektion: Inhaltskriterien

- Sind bestimmte Items als „Eisbrecher“ notwendig?
- Kann man auf bestimmte Items verzichten, weil genügend ähnliche vorhanden sind?
- Repräsentieren bestimmte Items ein Testmerkmal besonders prägnant trotz ungünstiger Itemkennwerte?
- Lassen sich bestimmte Items trotz guter Kennwerte nicht theoretisch-inhaltlich korrekt einordnen?
- Verletzen bestimmte Items ethische Normen?
- Sind die Items (besonders bei ja/nein-Antworten) ausbalanciert, d.h. gibt es etwa gleich viele „Ja“-Antworten wie „Nein“-Antworten?
- Sind die Items zumindest einigermaßen frei von Tendenzen zur sozialen Erwünschtheit?

Itemselektion: Statistik

Dieser Punkt untergliedert sich in die Betrachtung von

- Trennschärfe
- Itemschwierigkeit
- Streuung
- Evtl. zusätzlich: Dimensionalitätsüberprüfung

Definition:
Die Trennschärfe eines Items gibt an, wie gut das gesamte Testergebnis aufgrund dieses einzelnen Items vorhersagbar ist.
Sie ist ein Kennwert dafür, in welchem Ausmaß die Differenzierung der Versuchspersonen in Löser und Nicht-Löser durch das Item mit demjenigen durch die Skala als Ganzes übereinstimmt.

Um so höher die Trennschärfe, desto besser misst das Item das, was auch die Skala misst.

Das Kriterium *Trennschärfe I*

Faustregel: Wenn die Itemwerte und die Summenwerte weniger als 10 % gemeinsame Varianz haben, ist das Item ungeeignet. Die gemeinsame Varianz berechnet man durch Quadrierung der Korrelation, hier der Trennschärfe. Wenn man die Trennschärfe 0.32 quadriert, erhält man den Wert 0.1024, also 10,24 % gemeinsame Varianz. Das Quadrat von 0.31 liegt bereits darunter (unter der 10% - Grenze).

Also ist konventionsgemäß in der Regel 0.32 als untere Grenze akzeptabler Trennschärfe anzusehen.

Trennschärfe und Itemschwierigkeit

Theoretisch: Unabhängig von seiner Schwierigkeit (außer 0 und 100) könnte jedes Item eine Trennschärfe von 1.0 erreichen

Empirisch: zeigt sich jedoch eine umgekehrt u-förmige Beziehung zwischen Schwierigkeit und Trennschärfe, wobei mit mittlerer Schwierigkeit die höchste Trennschärfe einhergeht

Standardabweichung der Testwerte steigt mit den Trennschärfen

Das Kriterium *Trennschärfe IV*

Die Variation von Itemschwierigkeiten: führt zu einer Abnahme der Interkorrelationen zwischen den Items, damit zu einer Abnahme der Homogenität und zu einer Abnahme der Trennschärfe.

- Berechnung von Trennschärfen
- Inhaltliche Erläuterungen zu Trennschärfe und Schwierigkeit
- Berechnung von Trennschärfen mit SPSS
- Beispiel einer Trennschärfenanalyse

Trennschärfenanalyse

- Inhaltlich drückt eine Trennschärfe aus, wie gut ein Item eine Skala, die aus den restlichen Items gebildet wird, widerspiegelt
- Eine Trennschärfe stellt die korrigierte Korrelation (Part-whole-Korrektur) einer Aufgabe mit einer Skala dar

Part-whole-Korrektur:

- **Ohne part-whole-Korrektur kommt es zu einer Überschätzung der Trennschärfe, da das betreffende Item selbst Bestandteil der Skala ist**
- **Ohne part-whole-Korrektur ginge ein Teil der Skalenstreuung auf das entsprechende Item zurück, mit dem die Skala korreliert wird**
- **Je größer die Itemanzahl einer Skala ist, desto geringer sind die Auswirkungen der Korrektur auf die Trennschärfe, denn mit zunehmender Itemzahl wird der Beitrag eines einzelnen Items relativ zum Gesamtskalenwert geringer**
- **Je homogener eine Skala ist, desto weniger ändern sich die Trennschärfen durch eine part-whole- Korrektur**

Trennschärfenanalyse

	„Ich gehe gerne auf Parties“		∑ aller Extraversions-items
Justus	1		17
Peter	5		48
Bob	3		26

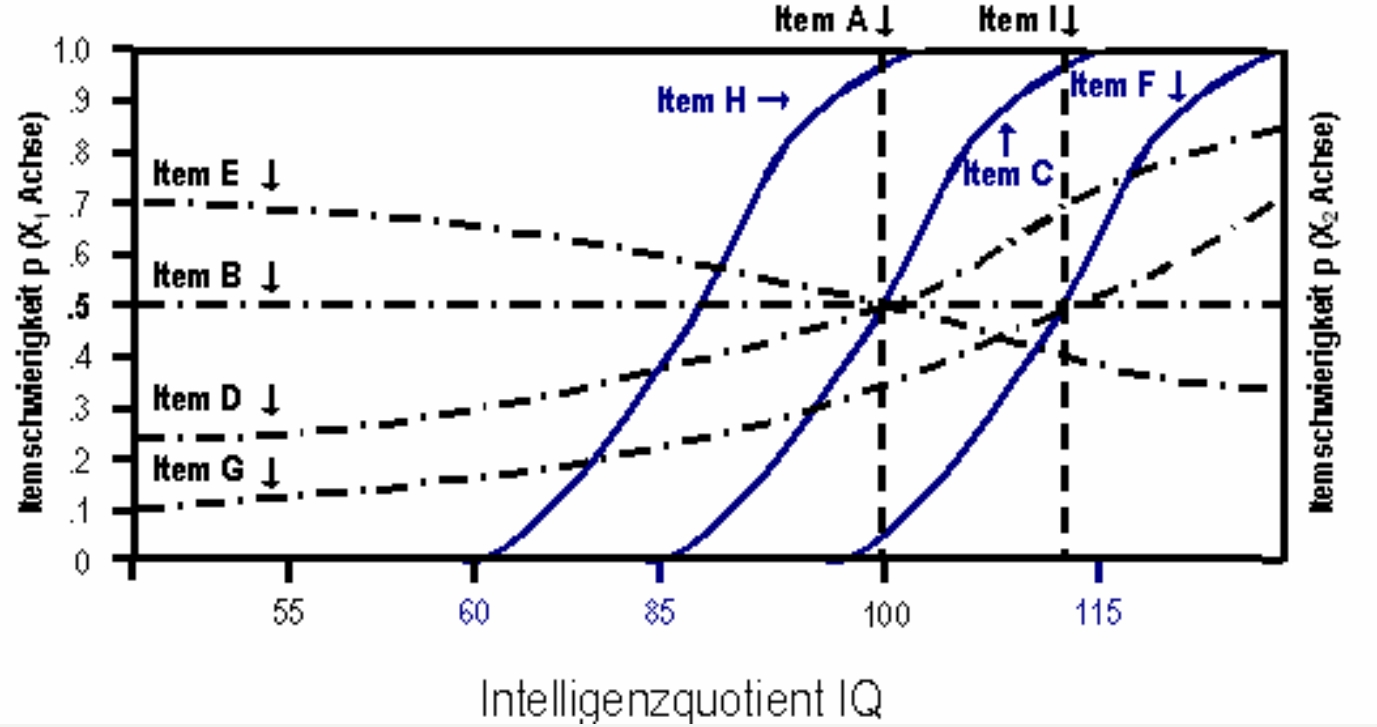
Trennschärfenanalyse

	„Ich gehe gerne auf Parties“	∑ aller Extraversions-items ohne Item 1	∑ aller Extraversions-items
Justus	1	16	17
Peter	5	43	48
Bob	3	23	26

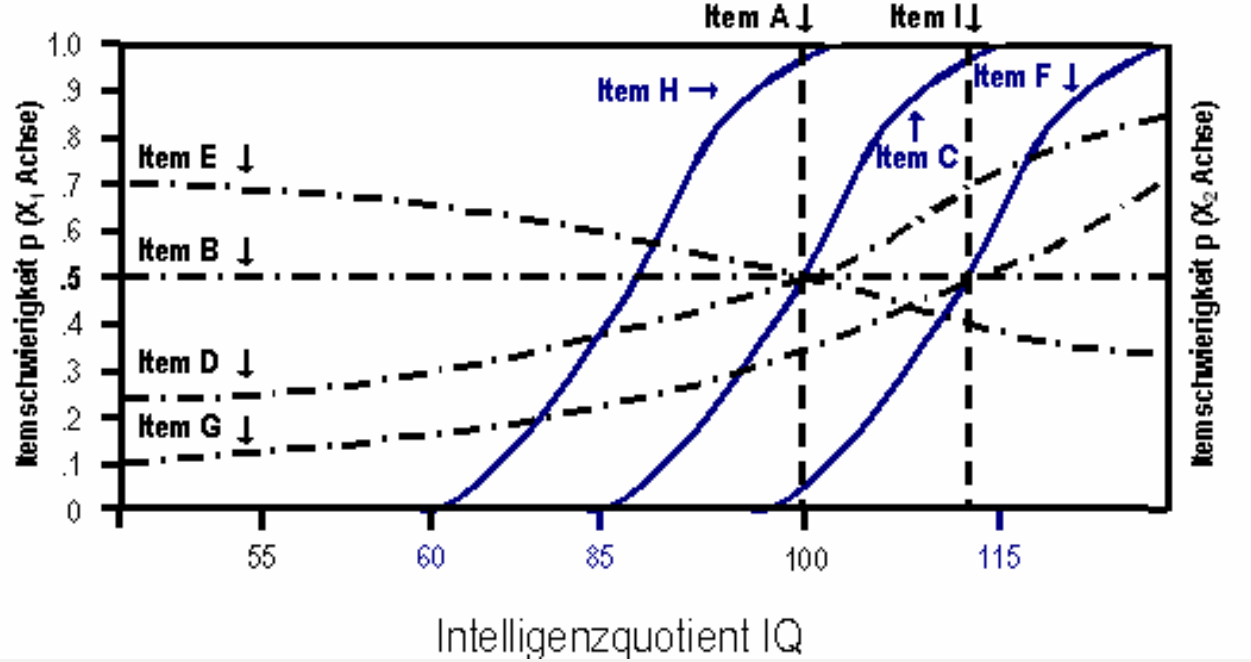
Zusammenhang zwischen Schwierigkeit und Trennschärfe

- Items mit mittlerer Schwierigkeit differenzieren am besten zwischen Probanden, die ein Item lösen (“Löser”), und Probanden, die ein Item nicht lösen (“Nicht-Löser”)
- Bei dichotomen Items ist die Itemstreuung rechnerisch vollkommen durch die Itemschwierigkeit determiniert
- Reichen die Itemschwierigkeiten bei intervallskalierten Items an den Rand der Antwortskala, spricht man von Boden- oder Deckeneffekten
- Beide Effekte haben zur Folge, dass zwischen Individuen mit verschiedenen Merkmalsausprägungen nicht mehr ausreichend differenziert werden kann

Kombination unterschiedlicher Itemschwierigkeiten mit unterschiedlichen Trennschärfen:

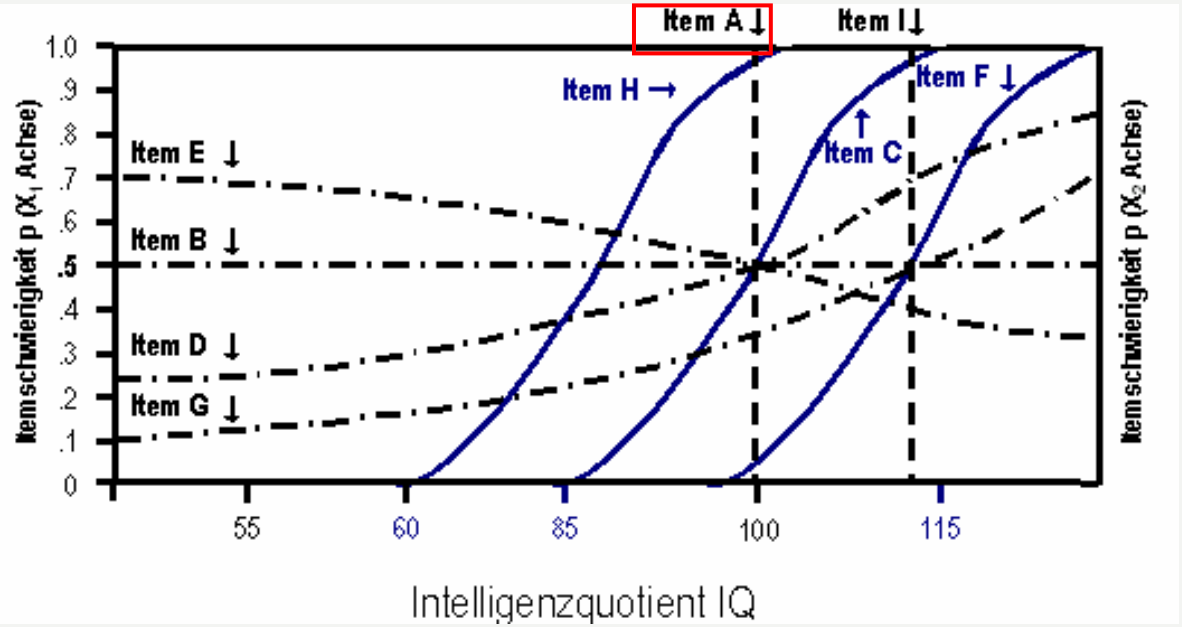


Trennschärfenanalyse



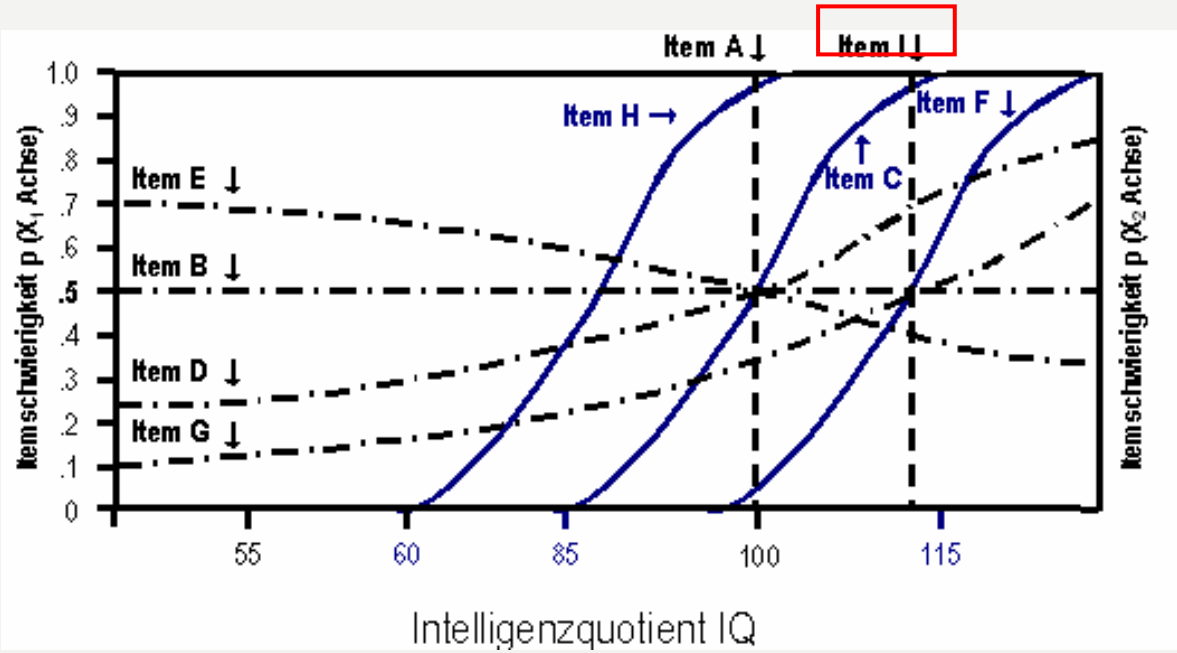
Je steiler der Anstieg der Item Characteristic Curves (ICC), desto größer ist die Trennschärfe

Trennschärfenanalyse



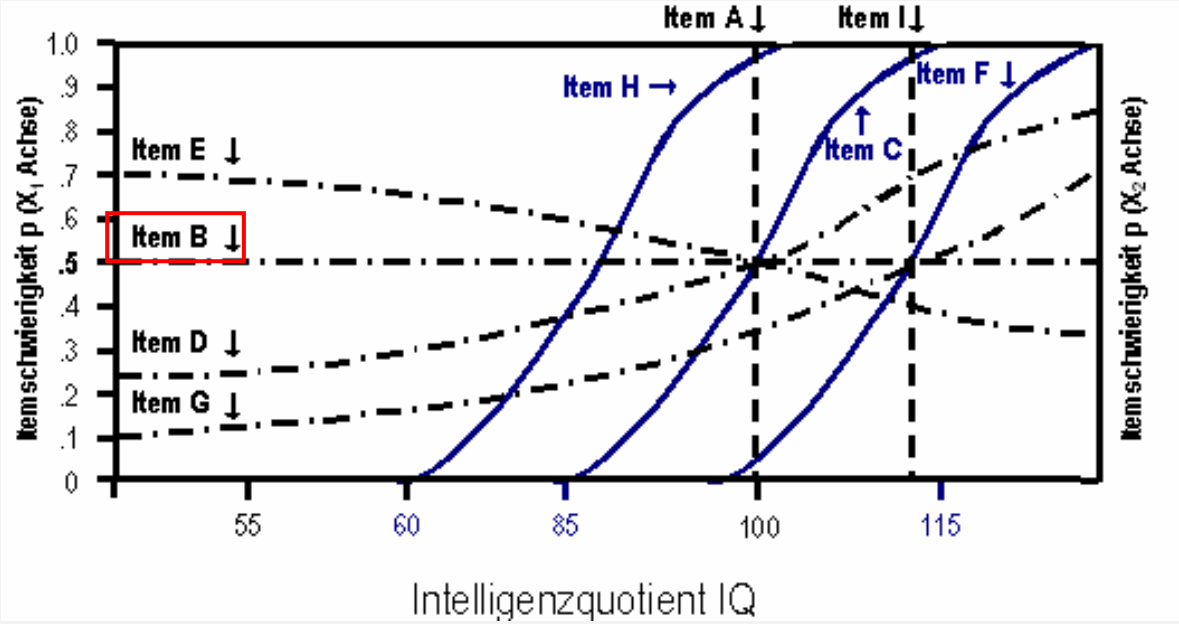
Item A ist ideal trennscharf ($p = .50$, $r_{it} = 1$).
Nur mit diesem Item alleine ließe sich entscheiden, ob ein Proband beispielsweise unter- oder überdurchschnittlich intelligent ist

Trennschärfenanalyse



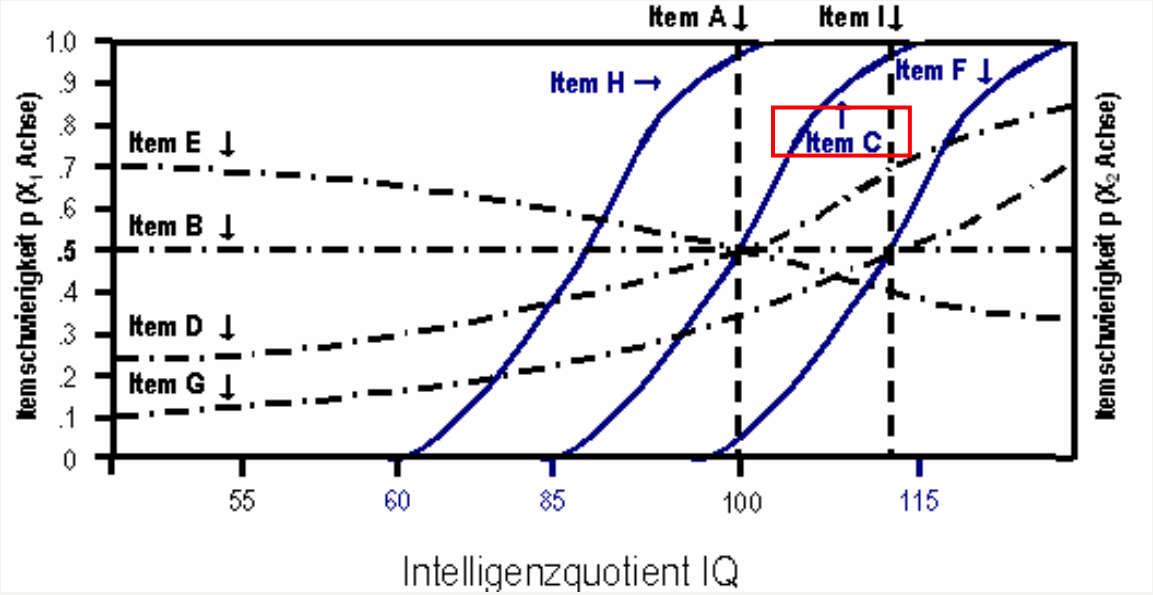
Mit Aufgabe I ($p = .20$, $r_{it} = 1$) ließe sich entscheiden, ob ein Proband zu den etwa 20 Prozent intelligentesten Probanden (IQ . 113) gehört oder nicht

Trennschärfenanalyse



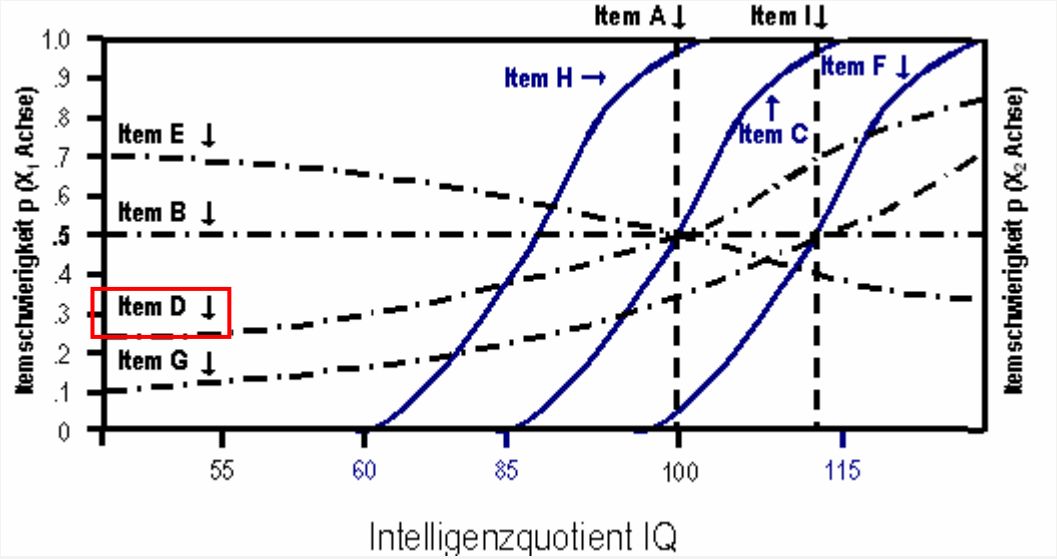
- Item B (p = .50, r_{it} = 1 = 0) dagegen ist vollkommen nutzlos, da es Intelligente von Nicht-Intelligenten nicht unterscheidet, obwohl es aufgrund seiner mittleren Schwierigkeit eigentlich gute Voraussetzungen besitzt

Trennschärfenanalyse



- Item C wird von keinem Probanden mit einem IQ unter 85 und von allen Probanden mit einem IQ über 115 richtig beantwortet (erkennbar durch das Auftreffen der ICC auf die X1-Achse bzw. X2-Achse), es hat also eine hohe Trennschärfe

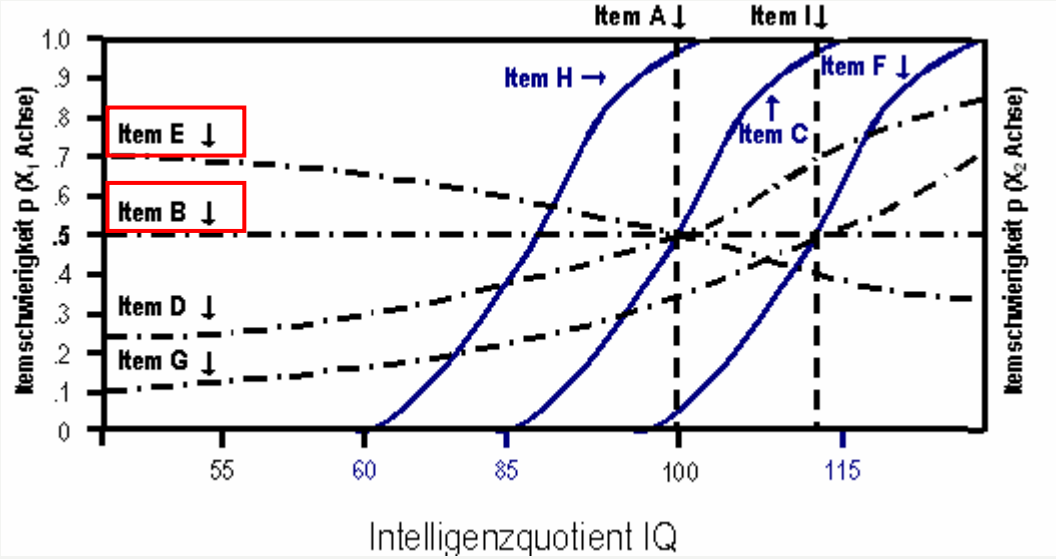
Trennschärfenanalyse



Das Item D ($p = .50$, $r_{it} = .30$) stellt den weitaus häufigsten Fall eines Items mit mittlerer Trennschärfe bei gleichzeitig geringer bis mittlerer Itemschwierigkeit dar

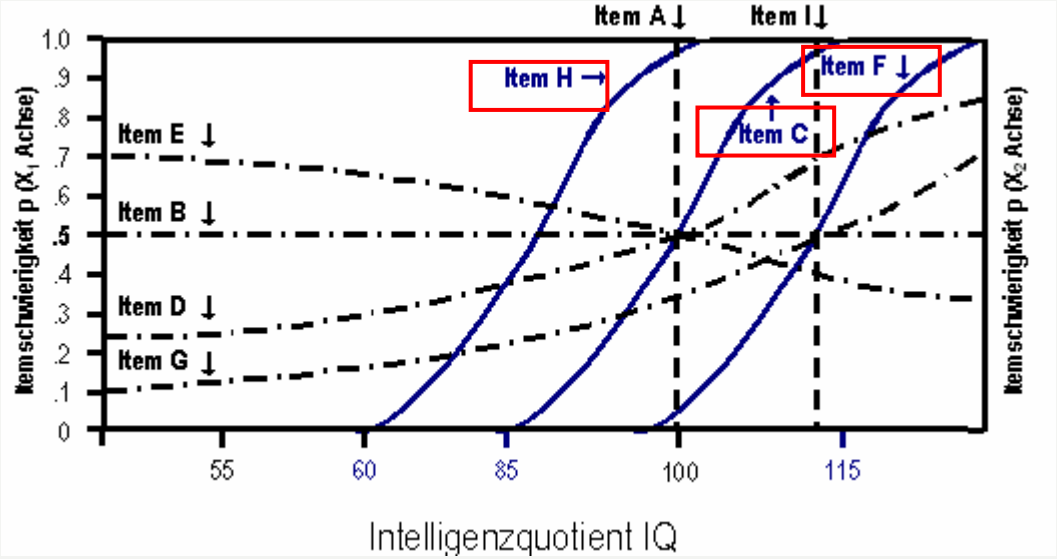
Mit Items dieser Art kann man eine Differenzierung entlang der gesamten Testskala erreichen

Trennschärfenanalyse



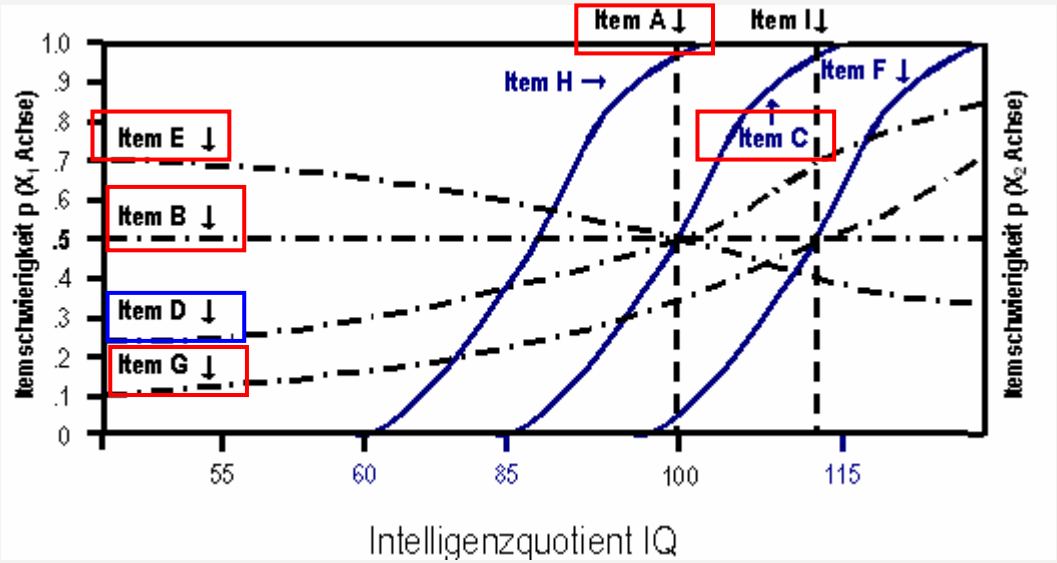
- Im Gegensatz zu allen bisherigen Items, wird Item E von den weniger intelligenten Probanden etwas häufiger gelöst als von den intelligenten; es hat folglich eine geringe und negative Trennschärfe
- Solche Items sind ebenso wie Item B für die Testkonstruktion unbrauchbar

Trennschärfenanalyse



- Die Items F und H differenzieren gut, aber nur in extremen Schwierigkeitsbereichen (IQ . 87 und 113)
- Die Items C, F und H haben zwar die gleiche Trennschärfe (gleicher Anstieg der ICC), aber unterschiedliche Schwierigkeit

Trennschärfenanalyse



- Die Items A, B, C, D und E haben die gleiche Schwierigkeit (ICC´s schneiden sich bei IQ = 100), aber unterschiedliche Trennschärfen (unterschiedlicher Anstieg der ICĆs)
- Item G hat eine mittlere Trennschärfe (flacher Anstieg der ICC) bei einer Schwierigkeit von p = .20 (p wie bei Item I)

Trennschärfenanalyse

- Insgesamt differenzieren Tests mit homogen mittelschweren Items am besten bei mittleren Merkmalsausprägungen
- Da bei mittlerer Itemschwierigkeit die Wahrscheinlichkeit für hohe Trennschärfen ansteigt, ist für solche Skalen auch eine höhere Reliabilität zu erwarten
- Um auch in Randbereichen eines Merkmalsbereichs zu differenzieren, muss die Skala auch extremere Schwierigkeitsbereiche mit Items abdecken
- Meist erreichen Items mit extremen Schwierigkeiten geringere Trennschärfen als mittelschwere Items. Dies reduziert die Itemhomogenität und daher sind für solche Skalen nicht ganz so hohe Reliabilitäten wie für Skalen mit ausschließlich mittelschweren Items zu erwarten

Berechnung von Trennschärfen mit SPSS