LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Vorlesung Testtheorien

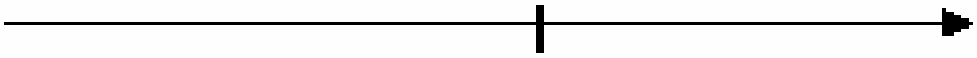
Dr. Tobias Constantin Haupt, MBA

Sommersemester 2007

Abbildung 6-1: *Veranschaulichung* einer Zuordnung von Personen- und Itemparameter (PP, IP) auf einer eindimensionalen Skala.

Drei Fälle:

1. Fall: Die Wahrscheinlichkeit, daß die Person j das Item i löst, ist *gleich 0.50.*



PP = IP

2. Fall: Die Wahrscheinlichkeit, daß die Person j das Item i löst, ist *größer als 0.50.*

IP < PP

3. Fall: Die Wahrscheinlichkeit, daß die Person j das Item i löst, ist *kleiner als 0.50.*

PP < IP

(Fisseni, 1990)

(Itemcharakteristiken, IC - Funktion):

Diese beschreibt die Beziehung zwischen einem latenten Merkmal (Personenparameter) und dem Reaktionsverhalten auf ein (dichotomes) Item in Form einer Wahrscheinlichkeitsaussage.

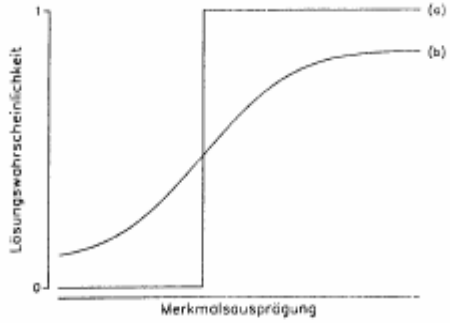
Itemcharakteristische Funktionen

Die verschiedenen Modelle der IRT unterscheiden sich darin, welche IC-Funktion angenommen wird. Grundsätzlich lassen sich folgende Typen von IC-Funktionen (Modelle) unterscheiden:

Deterministische Itemcharakteristiken:
Wenn davon ausgegangen wird, daß das Antwortverhalten der Versuchspersonen durch die Item- und Personenparameter vollständig bestimmt wird, d.h. die Lösungswahrscheinlichkeiten für die einzelnen Items je nach β und δ immer entweder Null oder Eins sind.

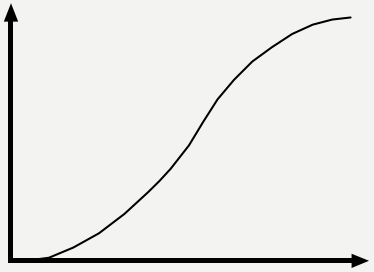
***Probabilistische Itemcharakteristiken*:**

Wenn eine stochastische Beziehung zwischen β, δ und der Lösungswahrscheinlichkeit angenommen wird, d.h. Lösungswahrscheinlichkeiten in allen Abstufungen zwischen Null und Eins auftreten können. Solche Funktionen sind in der Regel monoton steigend [d.h., je höher β (also die Fähigkeits-/Merkmalsausprägung einer Person), desto höher die Lösungswahrscheinlichkeit].

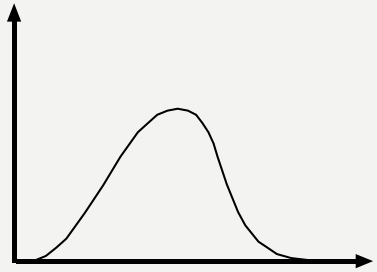


Zwei ***monotone*** Itemcharakteristiken, (a) deterministisch, (b) probabilistisch

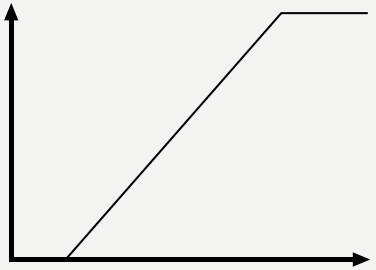
Arten von Item-Charakteristiken



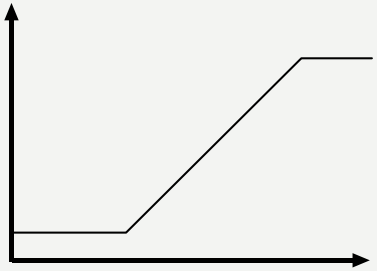
monotone Charakteristik



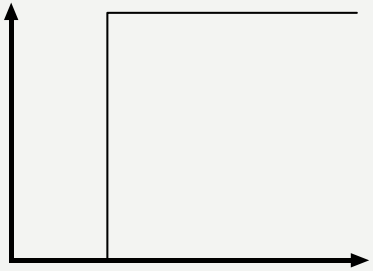
nicht-monotone Charakteristik



lineare Charakteristik



lineare Charakteristik mit Rate- und Fehler-wahrscheinlichkeit



Guttman-Charakteristik

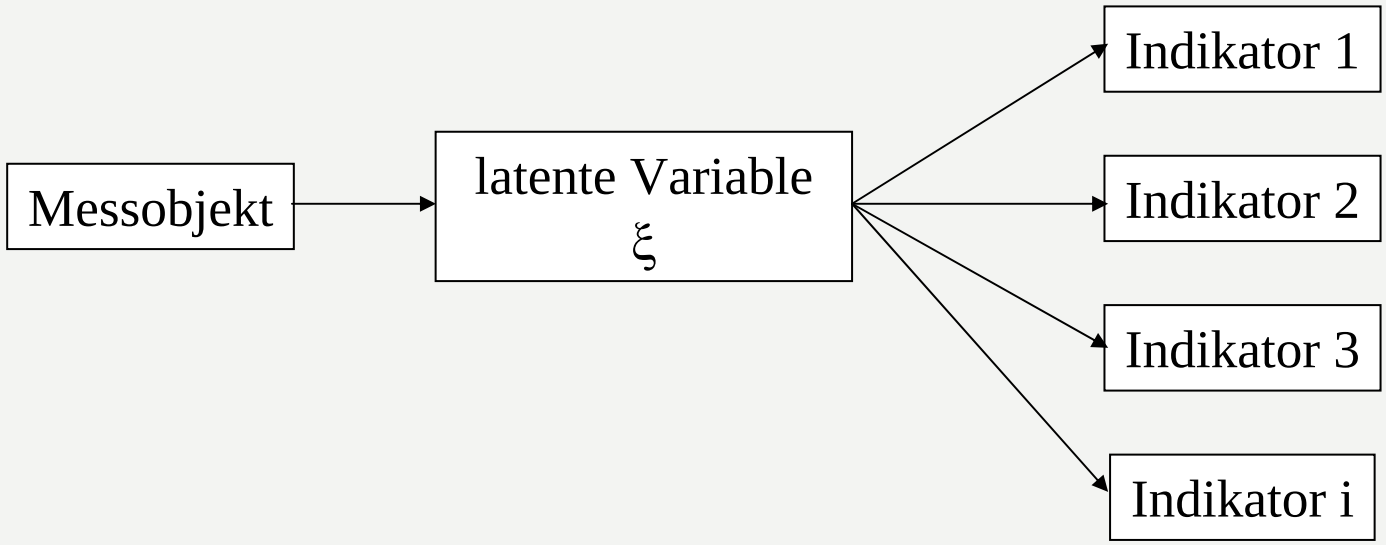
Grundannahmen der IRT I

Die jeweils gefragte Variable, auf der jedem Testobjekt ein Wert zugeordnet ist, ist eine ‚latente Variable'

- die nicht direkt zugänglich ist,
- für die Indikatoren existieren

Die jeweils gefragte Variable, auf der jedem Testobjekt ein Wert zugeordnet ist, ist eine ‚latente Variable'

- die nicht direkt zugänglich ist,
- für die Indikatoren existieren



lokale stochastische Unabhängigkeit

Fragestellung: Wie könnte man prinzipiell von mehreren manifesten Variablen auf eine dahinterliegende (die Ausprägungen der manifesten Variable verursachende) latente Variable schließen?

Antwort: dies ist dann der Fall, wenn die

1. Items „homogen“ bezüglich der latenten Variablen sind, d.h., wenn die manifesten Variablen miteinander korrelieren,

2. die manifesten Variablen (inhaltlich) Indikatoren der latenten Variablen sind und

3. die latente Variable als Ursachenfaktor (Indikator) für die Korrelation der manifesten Variablen untereinander verantwortlich ist

lokale stochastische Unabhängigkeit

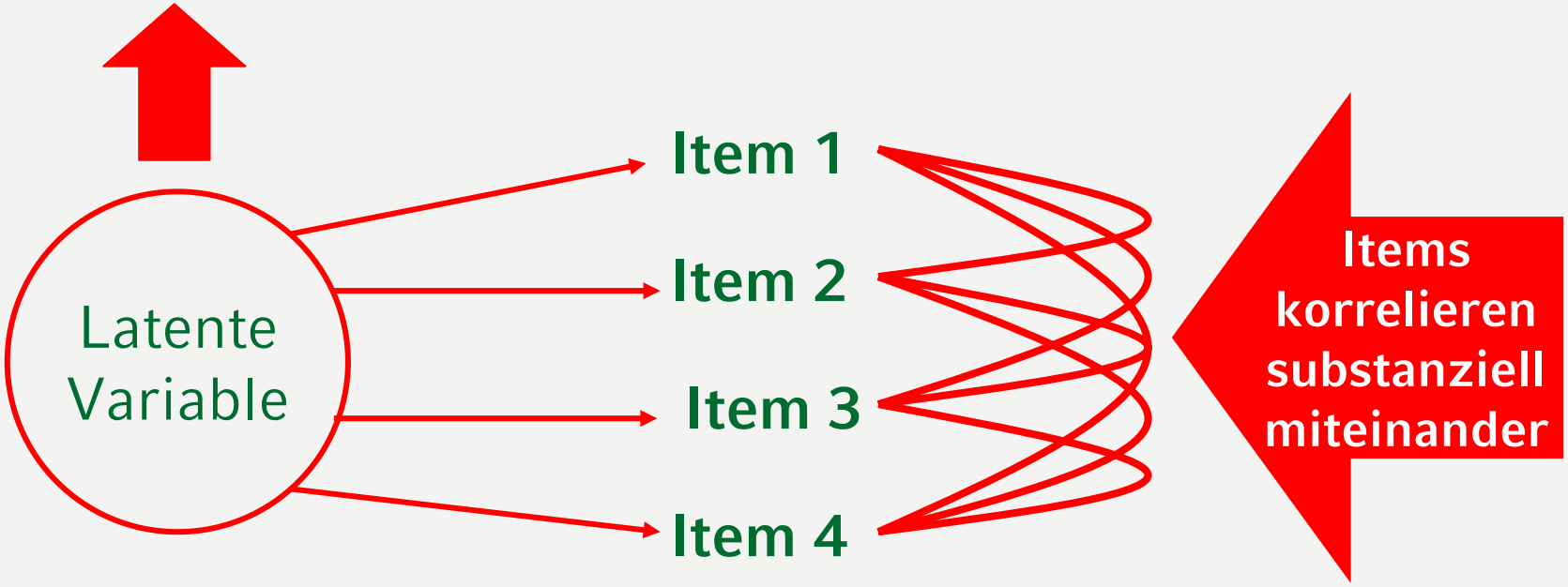
- Überprüfung: Itemhomogenität läge dann vor, wenn bei Herauspartialisierung des Einflusses von ξ aus der Korrelation zwischen den manifesten Variablen keine Korrelation mehr zwischen diesen bestünde

- Die Logik dabei ist, daß wenn nur die latente Merkmalsausprägung die Korrelation zweier Items auf einer Stufe verschwinden läßt (vgl. lokale stochachstische Unabhängigkeit), dann muß dies unabhängig von der Stichprobe sein! Oder anders herum: Ursache der Korrelation der manifesten Variablen ist dann einzig und allein die latente Variable.

Lokale stochastische Unabhängigkeit

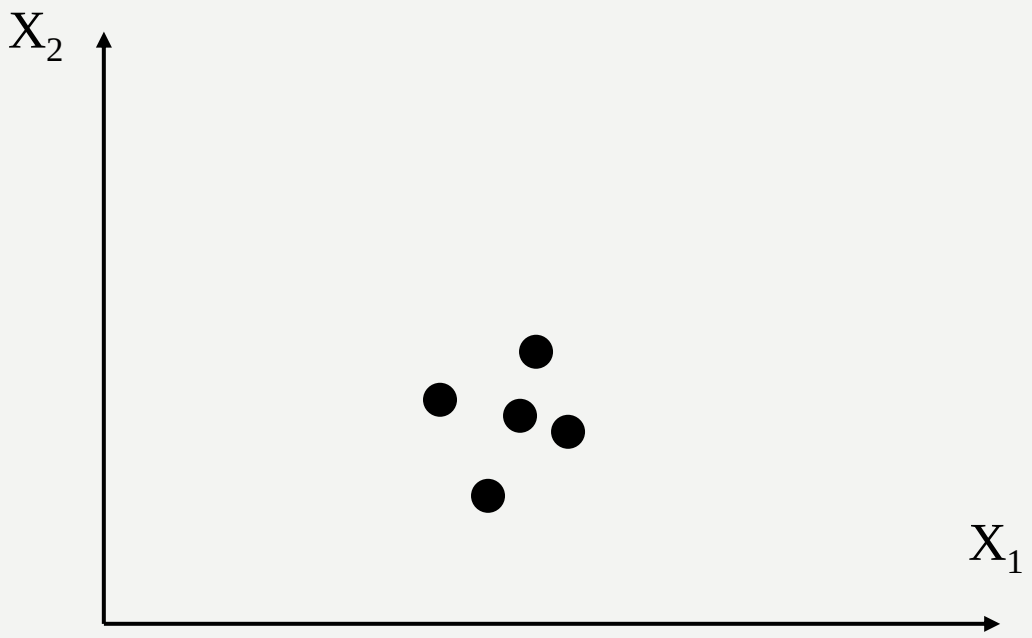
Konstanthaltung eines Wertes der latenten Variablen

→ Kein Zusammenhang zwischen den Items mehr bei Konstant-haltung der latenten Variablen auf einen Wert



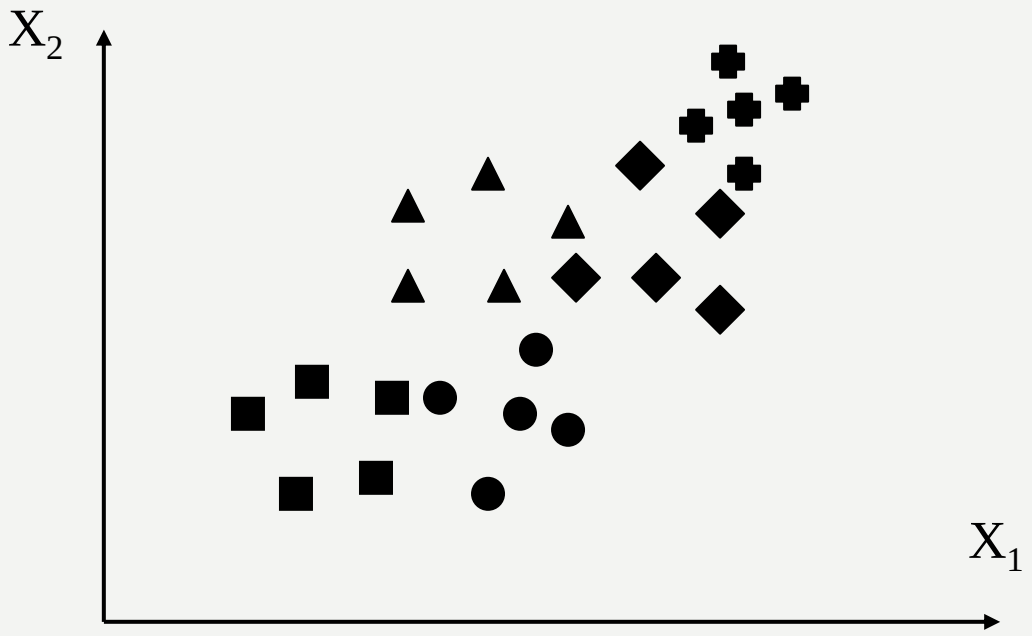
Lokalität von Unabhängigkeit

! *lokale* Unabhängigkeit schließt Korrelation in der *Population* nicht aus



Lokalität von Unabhängigkeit

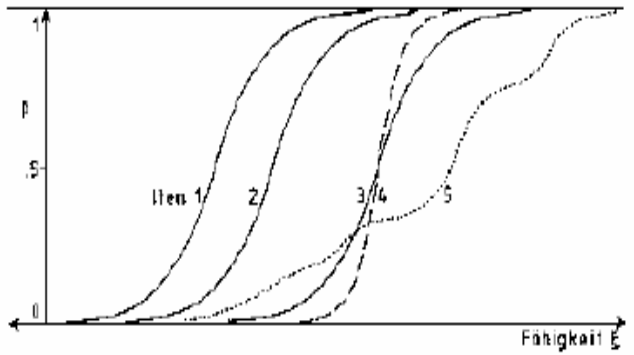
! *lokale* Unabhängigkeit schließt Korrelation in der *Population* nicht aus



Klassifikationskriterien

- ***nach Art der itemcharakteristischen Funktion***:
  - –deterministisch
  - –probabilistisch (z.B. linear monoton steigend, logistisch, etc.).

Abbildung 7.2: Itemcharakteristiken



Die Itemcharakteristiken der Items 1, 2, 3 entsprechen dem Rasch-Modell. Die Hinzunahme von Item 4 wäre im Birnbaum-Modell möglich. Item 5 hat eine unregelmäßig monoton steigende Itemcharakteristik.

Verschiedene Latent-Trait-Modelle lassen unterschiedliche Formen der Itemcharakteristik zu (siehe Abb. 7.2).

nach Variablenart der manifesten und latenten Variablen:

- *Latente Variablen: Können…*
  - *als kontinuierlich (unterschiedliche quantitative Ausprägungen) angenommen werden (Latent-Trait-Modelle), diese sind in der psychologischen Diagnostik am häufigsten, oder*
  - *nur qualitativ unterschiedliche Ausprägungen (liegt vor versus liegt nicht vor, also z. B. Persönlichkeitstypen) haben (Latent-Class-Modelle).*

Manifeste Variablen:

Können entweder

- dichotom (wie im dichotomen Rasch-Modell) sein oder
- abgestuft sein (Ratingskalen), z.B. der eindimensionale Spezialfall des polytomen Rasch-Modells.

nach Anzahl der Modellparameter:

Ob z.B.

- nur unterschiedliche Item- und Personenparameter angenommen werden müssen (z.B. Guttman-Modell oder dichotomes Rasch-Modell),
- ein variierender Itemdiskriminationsparameter notwendig ist (z.B. im Birnbaum-Modell) oder ob
- weitere Parameter verwendet werden

Allgemein

- Annahme eines **latenten Kontinuums** (Fähigkeit, Eigenschaft) ξ
- Jede Person v weist auf diesem eine bestimmte **Ausprägung** ξ_v auf.

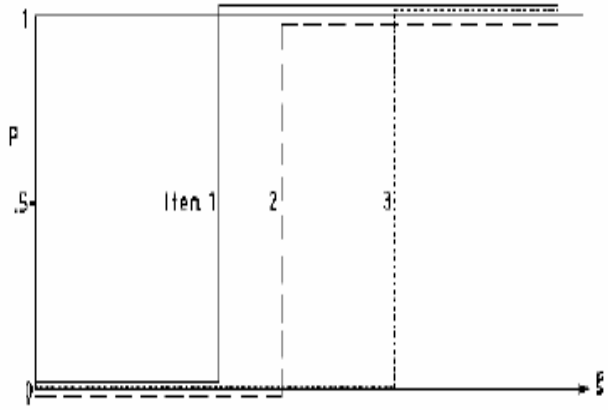


Abbildung 7.1: Guttman-Skala mit drei Items. Für jedes Item steigt an einer bestimmten Stelle des Merkmalskontinuums ξ die Lösungswahrscheinlichkeit p von Null auf Eins.

Es könnte einen **kritischen Wert** auf ξ geben, ab dem ein Item gelöst wird. → Grundgedanke der **Guttman**-Skala

Allen Latent-Trait-Modellen gemeinsam:

- Latentes Kontinuum
- Itemcharakteristik
- Lokale stochastische Unabhängigkeit

Unterschiede

- Form der Itemcharakteristik
- Folgerungen daraus
  (z.B. dichotom oder mehrkategorial)

Modellannahmen

Personparameter:

- Fähigkeit einer Person, ein bestimmtes Item zu lösen.
- Sie läßt sich durch einen Wert auf einer eindimensionalen Skala repräsentieren.

Itemparameter:

- Jede Itemschwierigkeit läßt sich durch einen Messwert auf einer eindimensionalen Skala repräsentieren.
- **Person- und Itemparameter** lassen sich **gemeinsam auf einer eindimensionalen Skala** abbilden.
- (PP >, < oder = IP ?)

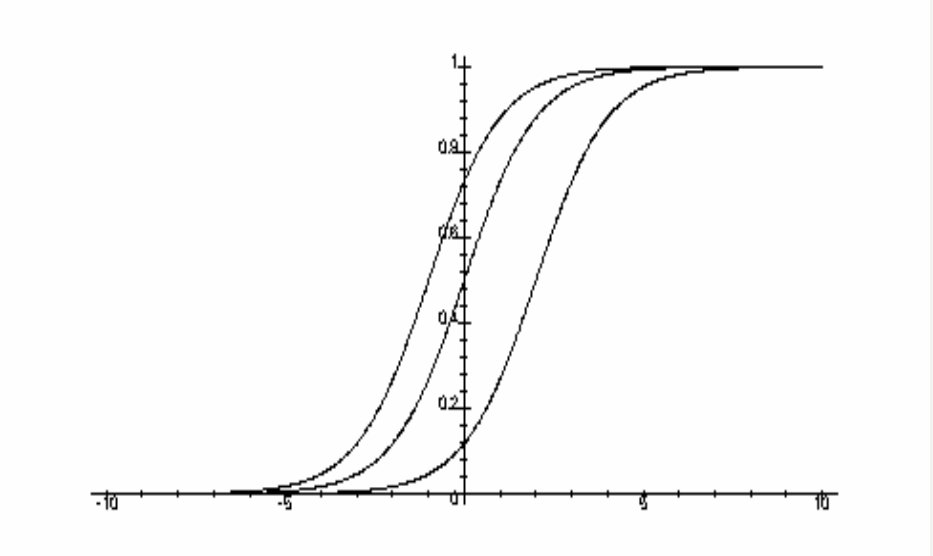
Das Rasch-Modell

Der Zusammenhang zwischen der Lösung eines Items und den beiden Parametern ist ***probabilistisch***:

„In Abhängigkeit von der Höhe von Item- und Personparameter läßt sich dem Ereignis **„Item wird gelöst“** ein **Wahrscheinlichkeitswert** zuordnen.“

Das Rasch-Modell

Diese Annahmen über Item- und Personparameter sollen in einer Wahrscheinlichkeitsfunktion abgebildet werden:



Einordnung:

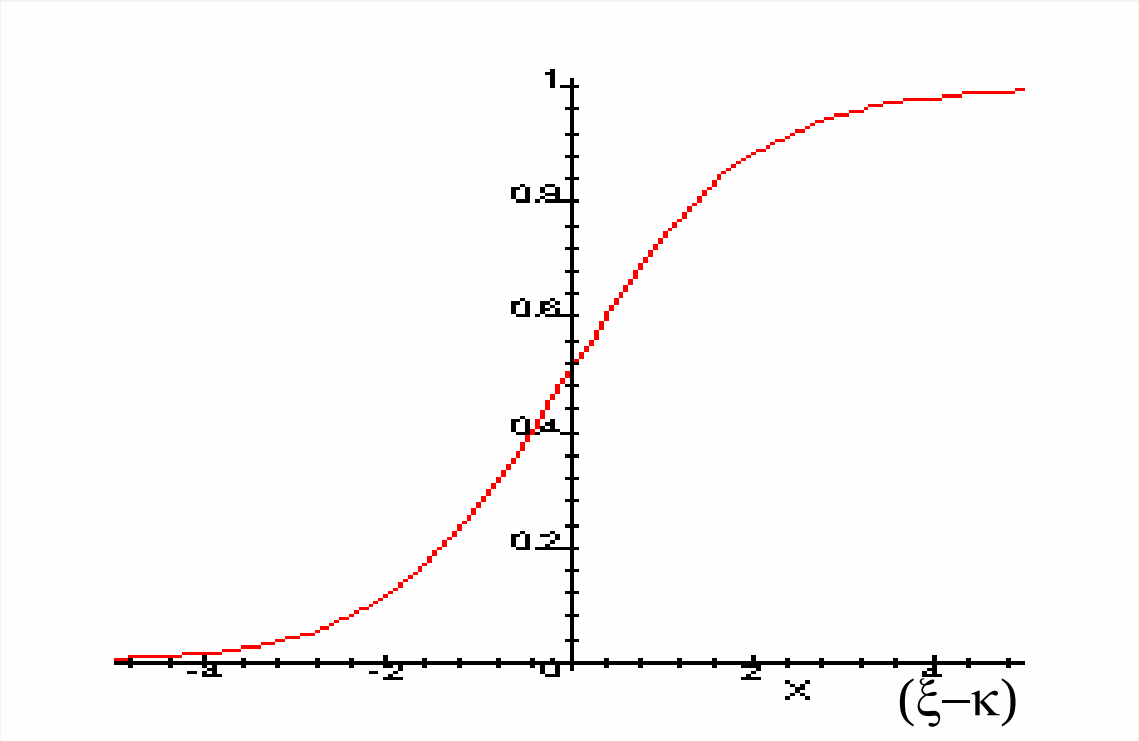
- Das dichotome Rasch-Modell ist ein probabilistisches Modell, welches kontinuierliche latente Variablen annimmt, sowie bei dichotomen manifesten Variablen (Alternativantworten) angewendet wird. Die zugrundeliegende IC-Funktion ist logistisch.

Dichotomes Rasch-Modell:

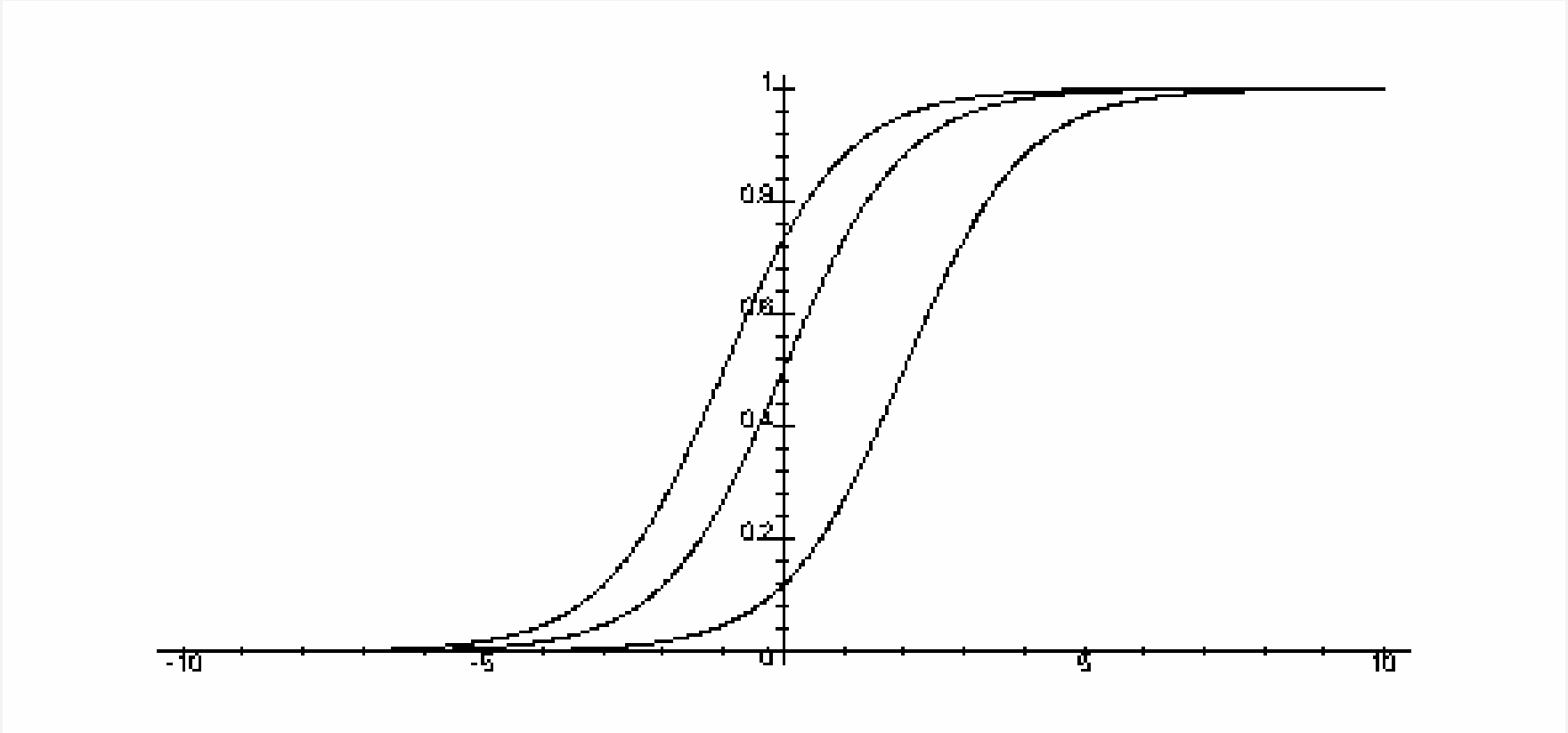
- Probabilistisch-logistisches dichotomes Latent-Trait-Modell mit invarianten Diskriminationsparametern.

Rasch-Modell: itemcharakteristische Funktion

$$P(+ \mid \xi, \kappa) = \frac{e^{(\xi - \kappa)}}{1 + e^{(\xi - \kappa)}}.$$

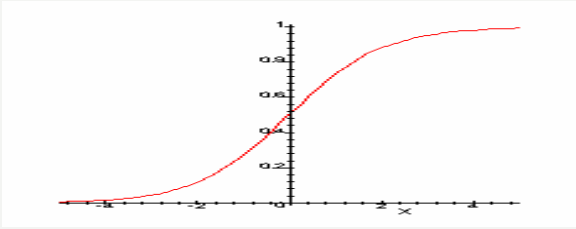


mehrere charakteristische Funktionen



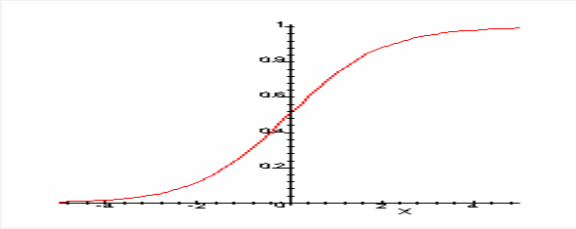
Modellgleichung und logistische IC-Funktion:

- Die Lösungswahrscheinlichkeit einer best. Person für ein best. Item p (x) wird im Rasch-Modell allein durch die Ausprägungen vom Fähigkeitsparameter β und vom Itemschwierigkeitsparameter δ bestimmt.
- Der Zusammenhang zwischen Parametern und Lösungswahrscheinlichkeit soll nun durch die sog. logistische Funktion festgelegt sein, welche die Eigenschaft hat, daß im Mittelbereich (dort, wo β und δ gleich sind) nahezu Linearität zwischen Fähigkeit und Lösungswahrscheinlichkeit besteht, während sich die Lösungswahrscheinlichkeiten im oberen und unteren Fähigkeitsbereich asymptotisch den Grenzwerten 0 und 1 nähern.



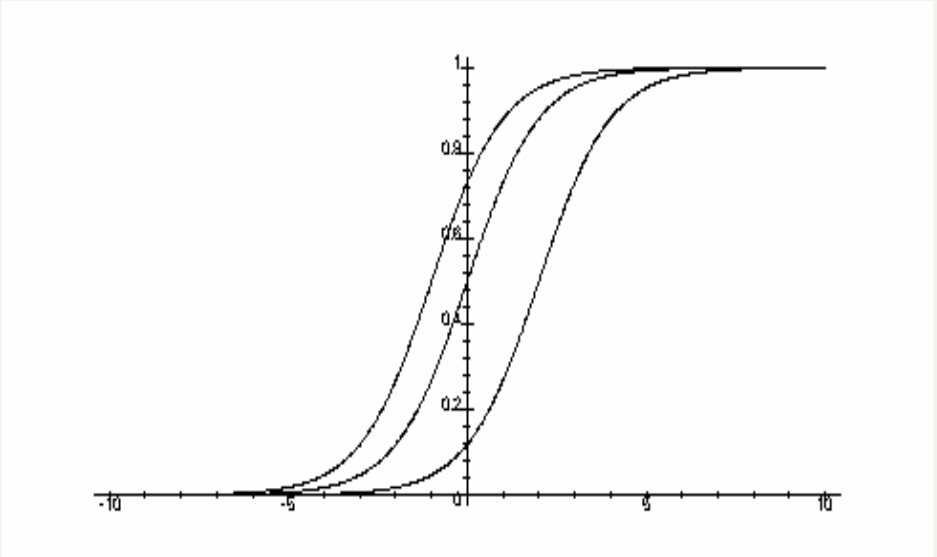
Änderungen von p(x) in Abhängigkeit von β und δ:

- Angenommen, Fähigkeit (β) und Itemschwierigkeit (δ) sind gleich groß, dann beträgt die Lösungswahrscheinlichkeit dieses Items 50%. An dieser Stelle hat die logistische Funktion ihren Wendepunkt.
- Je mehr die Fähigkeit die Itemschwierigkeit übersteigt, d.h., je positiver die Differenz (β-δ) wird, desto größer wird die Lösungswahrscheinlichkeit (wobei sie jedoch bei geringeren Differenzen schneller steigt).



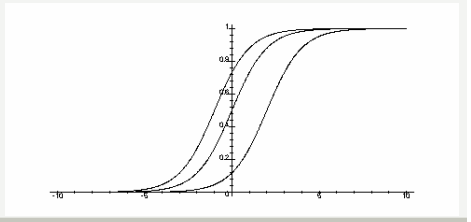
Eigenschaften einer Raschmodellkonformen-Skala:

Angenommen ein Itemsatz entspräche (was ja bislang noch nicht nachgewiesen ist) den Annahmen des Rasch-Modells. Dann ergeben sich bei der Anwendung solcher Skalen vier vorteilhafte Modelleigenschaften:



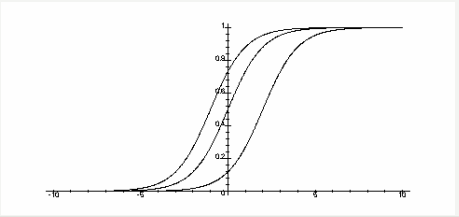
I. Itemhomogenität:

- Es werden nur itemcharakteristische Kurven zugelassen, die sich nicht schneiden, d.h., daß alle Items den gleichen Verlauf der Lösungswahrscheinlichkeiten zeigen (in diesem Sinne sind sie homogen).
- Sie unterscheiden sich lediglich darin, daß sie an unterschiedlichen Stellen des Item – Personenparameter - Kontinuums laufen (je höher δ, desto weiter rechts). Das bedeutet also, daß die IC – Kurven parallel entlang der x – Achse verschoben sind.
- Dabei gilt für jedes Item: die Wahrscheinlichkeit, dieses Item zu lösen, ist für „tüchtigere" Personen immer größer als für weniger tüchtige. Items, die nicht homogen sind, werden bei der Testkonstruktion eliminiert.



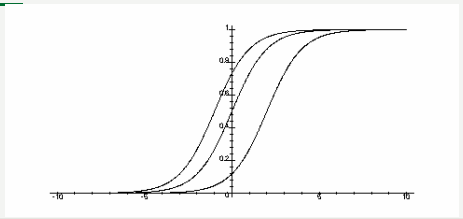
II. Erschöpfende Statistiken:

- **Wenn** Items lokal stochastisch unabhängig voneinander sind, d.h. wenn die Wahrscheinlichkeit, ein Item zu lösen nicht von der Wahrscheinlichkeit abhängt, ein anderes Item zu lösen, sondern ausschließlich von Fähigkeit und Itemschwierigkeit (s. o.),
- dann liefert allein die **Anzahl** der gelösten Items (unabhängig davon, welche Items, bzw. welche Itemteilmengen gelöst worden sind) eine „erschöpfende Statistik“ für die Fähigkeit einer Person.
- Ebenso liefern die **Anzahl** der Versuchspersonen (unabhängig davon welche Versuchspersonen das Item bearbeiten) eine erschöpfende Statistik für den Itemparameter.



III. Spezifische Objektivität (Teilgruppenkonstanz):

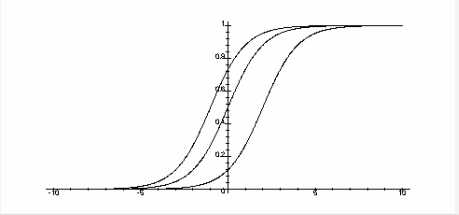
- Innerhalb einer Population, für die Modellkonformität festgestellt worden ist, fallen für einen Probanden (und auch bei Probandenvergleichen) sowohl Item- als auch Personenparameter immer gleich aus, gleichgültig, welche Merkmalsausprägung der Proband hat und unabhängig von den Items, die bearbeitet worden sind
- Diese Eigenschaft steht im Gegensatz zur KTT, wo zwei Versuchspersonen ihre Rangplätze vertauschen können, wenn man ihre Leistung nach Teilmengen der Items beurteilt.
- *Ergo: es besteht Unabhängigkeit beim Vergleich zweier Personen von dem Instrument, anhand dessen der Vergleich vorgenommen wurde!*



IV. SP-Unabhängigkeit der Parameterschätzungen (Separierbarkeit der Parameter):

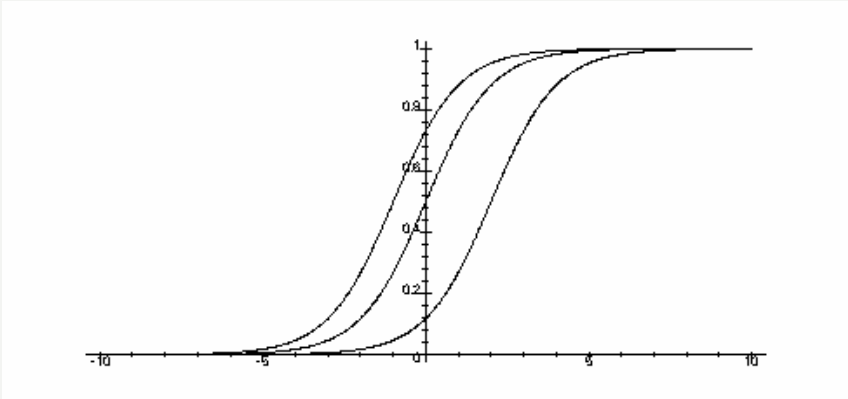
- Itemparameter können unabhängig von der Kenntnis der Personenparameter und Personenparameter unabhängig von Itemparametern geschätzt werden.

- Dies hat den Vorteil, daß man keine Verteilungsannahmen über unbekannte Parameter machen muß.



Empirische Modelltests:

Die Modellkonformität kann empirisch geprüft werden, indem man schaut, ob die oben angeführten Eigenschaften des Rasch-Modells zutreffen.



Ausgangsgleichung der Rasch-Skalierung

Die logistische Funktion L(x):

- Die Differenz zwischen Item- und Person-parameter wird als Exponent **x** eingesetzt.
- Statt PP und IP verwendet man die Symbole β (für a***b*** ility) und δ (für ***d*** ifficulty).

Das **Problem** ist jetzt nur:

- Zu Beginn der Testkonstruktion sind weder **Schwierigkeit** noch **Fähigkeit** bekannt und müssen **geschätzt** werden.
- Schwierigkeitsindizes *p* wie in der KTT:

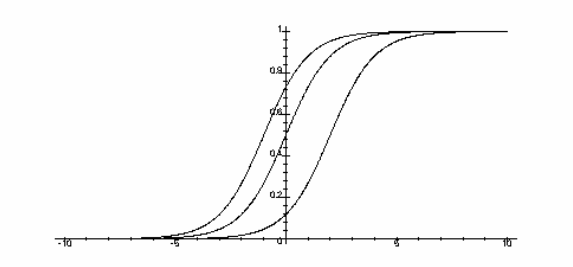
  → Wahrscheinlichkeit zur Itemlösung bei bekanntem Item- und Personparameter

Schritte bei der Rasch-Skalierung:

- Erstellung einer Matrix von **Schwierigkeit**sindizes
- Transformation in eine **Logit-Matrix**
- **Schätzung von Item- und Person-Parameter** aus der Logit-Matrix
- Reproduktion der Ausgangsmatrix als **Modelltest**

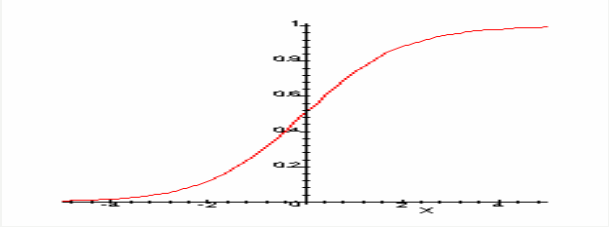
Die **Differenzierungsfähigkeit/Trennschärfe** von Items…

- …ist dort am größten, wo die logistische Funktion, bzw. die Lösungswahrscheinlichkeit die stärkste Steigung aufweist (Maximum der Iteminformationsfunktion).
- Die stärkste Steigung liegt am Wendepunkt vor, also dort, wo Item- und Personenparameter identisch sind, die Lösungswahrscheinlichkeit also 50 % beträgt.



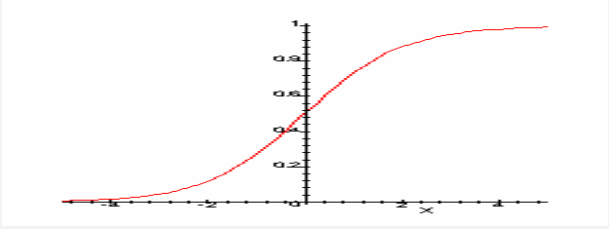
Adaptives Testen

- Das ökonomischste Testlängen-Informationsgewinn-Verhältnis würde sich nach der IRT jedoch dann ergeben, wenn δ und β identisch sind, man einer Person also nur Items vorlegen würde, deren Schwierigkeit dem Personenparameter möglichst ähnlich sein sollte.
- Dies ist die Idee des adaptiven Testens, wobei angepaßte Items anhand von Verzweigungen vorgelegt werden (meist per Computer).
  - [Möglich wird eine selektive Itemauswahl aufgrund von erschöpfenden Statistiken und spezifischer Objektivität]



Das **adaptive Testen** kann dabei untergliedert werden in das

- **„tailored testing“,** welches meist computergestützt durchgeführt wird und bei dem jedes Item in Abhängigkeit von der Beantwortung vorheriger Items ermittelt wird, ob es voraussichtlich optimal „paßt“ (d. h. über den Fähigkeitsparameter der Personinformiert) und dem
- **„branched testing“,** z. B. im AID von Kubinger & Wurst realisiert, wo auf Papier und Bleistift – Basis kleinere Itemgruppen/Subtests vorgegeben werden und dann in Abhängigkeit der Antworten für diese Itemgruppe/Subtest die beste nächste Itemgruppe ermittelt wird



Kritik der IRT - Pro

- ***Allgemein***:
  Verbesserte messtheoretische Eigenschaften.

- ***Empirische Prüfbarkeit der Modelleigenschaften***:
  Die Axiome der KTT können im Gegensatz zu den IRT-Modellen nicht empirisch auf Adäquatheit geprüft werden.

- ***Stichprobenunabhängigkeit***:
  Während in der KTT Aussagen über die Fähigkeit von Personen immer auf Items und ihre Lösungshäufigkeit in einer best. SP bezogen werden, sind in der IRT beide Parameter als getrennte und während der Konstruktion separierbare Größen konzipiert.

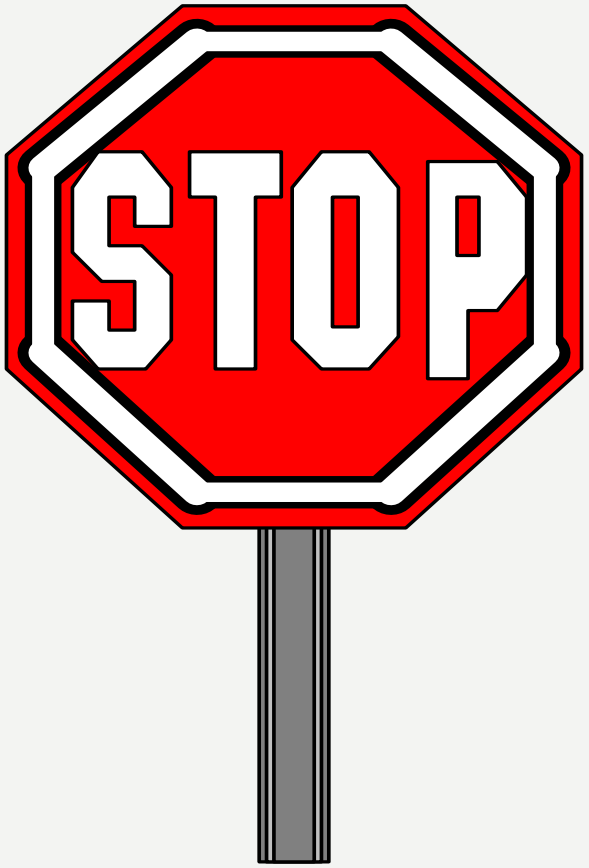
Kritik der IRT - Pro

- ***Intervallskalenniveau***:
  Liegt bei der IRT gesichert vor, während dies bei der KTT oft fraglich ist.
- ***Möglichkeit zum adaptiven Testen***:
  Ermöglicht die Durchführung ökonomischerer Tests; außerdem vermutlich motiviertere Probanden
- ***Anwendungsgebiet***:
  Erfolg versprechend sind Testkonstruktionen nach der IRT insbesondere da, wo Merkmale bereits theoretisch präzise definiert sind, und damit die zeitaufwendige Suche nach modellkonformen Items entfällt

Kritik der IRT - Contra

- **enormer Testkonstruktions – Mehraufwand**
- **die Art der Testkonstruktion schränkt den Testgegenstand ein: Schmaler Merkmalsbereich**: Items von Rasch-homogenen Skalen können einander sehr ähnlich werden
- **schwierige Reliabilitäts- und Validitätsüberprüfung**: Die Überprüfung der klassischen Testgütekriterien bereitet den probabilistischen Tests Schwierigkeiten; hinsichtlich der Validität drohe Gefahr „mit Kanonen auf Spatzen zu schießen“.
- **Wenige Konstruktionen bislang**: Tatsächlich sind bislang gemessen an theoretischen Veröffentlichungen und anerkennenden Worten über die IRT nur sehr wenige Tests konstruiert worden, die den Anforderungen der IRT genügen.

Warnhinweis für tiefergehend Interessierte zur IRT



Die Symbole für Variablen und Parameter werden nicht einheitlich, sondern **höchst unterschiedlich** benutzt!