

19. Testtheorie: Probabilistische Testtheorie PTT - Grundideen

<p>Grundgedanke: Item-Response-Theory IRT</p>	<p>Wie wahrscheinlich ist es, probabilistischer Zusammenhang ! dass sich eine bestimmte Merkmalsausprägung in einer bestimmten Reaktion auf ein Item zeigt ?</p>																		
<p>vgl. KTT deterministischer Zusammenhang: Merkmalsausprägung zeigt sich direkt in der Reaktion auf das Item</p>	<p>„Verantwortlich dafür, ob ein Proband ein Item löst (+) oder nicht (-), sind dabei sein Personenparameter θ (Theta) und der Itemparameter σ (Sigma)“ <i>(Markus Bühner, Einführung in die Testkonstruktion)</i></p> <p>→ Je weiter die die Personenfähigkeit die Itemschwierigkeit übersteigt, desto höher ist die Wahrscheinlichkeit, dass die Person das Item löst</p>																		
<p>Lösungswahrscheinlichkeit: Personenparameter + Itemparameter</p> <p>+ Trennschärfeparameter, Ratewahrscheinlichkeit, Schwellenparameter (je nach Messmodell)</p>	<p>Lösungswahrscheinlichkeit ergibt sich nicht aus den Varianzen, sondern aus dem Antwortmuster - in Abhängigkeit von „Personenstärke“ und „Aufgabenschwierigkeit“</p> <p><i>z.B. je intelligenter die Vpn, desto wahrscheinlicher löst sie eine Aufgabe mit einem bestimmten Schwierigkeitsgrad</i></p> <p>vgl KTT: <i>Die tatsächliche Personenfähigkeit kann eigentlich nur über unendlich viele Messwiederholungen ermittelt werden → wenn sich der Messfehler (irgendwann) auf 0 einpendelt</i></p>																		
<p>im Gegensatz zur KTT...</p>	<p>werden in der PTT die Annahmen / Voraussetzungen zur Anwendung eines bestimmten Messmodells in einem Modelltest überprüft</p> <p>z.B. mit dem <u>Rasch-Modell</u></p> <p>→ Intervallskalenniveau ? → Eindimensionalität ?</p> <div style="border: 1px solid yellow; padding: 5px; width: fit-content; margin-left: auto; margin-right: auto;"> <p>macht die KTT nicht – die Tests können aber auch im Rahmen der KTT eingesetzt werden !</p> </div>																		
<p>Messmodelle der PTT</p> <p><i>Beispiele</i></p>	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 15%;">dichotome Items</td> <td style="width: 25%; text-align: center;">Dichotomes Rasch-Modell</td> <td style="width: 25%; text-align: center;">Birnbaum-Modell (2PL-Modell)</td> <td style="width: 35%; text-align: center;">3PL-Modell</td> </tr> <tr> <td>Parameter</td> <td>*Itemschwierigkeit *Personenfähigkeit</td> <td>*Itemschwierigkeit *Personenfähigkeit *Trennschärfe</td> <td>*Itemschwierigkeit *Personenfähigkeit *Trennschärfe *Ratewahrscheinlichkeit</td> </tr> <tr> <td>ordinale Items</td> <td style="text-align: center;">Ordinales Rasch-Modell</td> <td style="text-align: center;">Ratingskalen-Modell</td> <td style="text-align: center;">Äquidistanz-Modell</td> </tr> <tr> <td>Parameter</td> <td>*Itemschwierigkeit *Personenfähigkeit *Schwellenparameter → Kategoriewahrscheinlichkeit</td> <td>*Itemschwierigkeit *Personenfähigkeit *gleiche Abstände zwischen aufeinanderfolgenden Schwellen über alle Items</td> <td>*Itemschwierigkeit *Personenfähigkeit *gleiche Abstände zwischen zwei Schwellen innerhalb eines Items, aber nicht über alle Items gleich groß</td> </tr> </table> <p style="text-align: center;">+ ordinales Mixed-Rasch-Modell zur Identifizierung von Mittel- und Extrem(an)Kreuzern</p>			dichotome Items	Dichotomes Rasch-Modell	Birnbaum-Modell (2PL-Modell)	3PL-Modell	Parameter	*Itemschwierigkeit *Personenfähigkeit	*Itemschwierigkeit *Personenfähigkeit *Trennschärfe	*Itemschwierigkeit *Personenfähigkeit *Trennschärfe *Ratewahrscheinlichkeit	ordinale Items	Ordinales Rasch-Modell	Ratingskalen-Modell	Äquidistanz-Modell	Parameter	*Itemschwierigkeit *Personenfähigkeit *Schwellenparameter → Kategoriewahrscheinlichkeit	*Itemschwierigkeit *Personenfähigkeit *gleiche Abstände zwischen aufeinanderfolgenden Schwellen über alle Items	*Itemschwierigkeit *Personenfähigkeit *gleiche Abstände zwischen zwei Schwellen innerhalb eines Items, aber nicht über alle Items gleich groß
dichotome Items	Dichotomes Rasch-Modell	Birnbaum-Modell (2PL-Modell)	3PL-Modell																
Parameter	*Itemschwierigkeit *Personenfähigkeit	*Itemschwierigkeit *Personenfähigkeit *Trennschärfe	*Itemschwierigkeit *Personenfähigkeit *Trennschärfe *Ratewahrscheinlichkeit																
ordinale Items	Ordinales Rasch-Modell	Ratingskalen-Modell	Äquidistanz-Modell																
Parameter	*Itemschwierigkeit *Personenfähigkeit *Schwellenparameter → Kategoriewahrscheinlichkeit	*Itemschwierigkeit *Personenfähigkeit *gleiche Abstände zwischen aufeinanderfolgenden Schwellen über alle Items	*Itemschwierigkeit *Personenfähigkeit *gleiche Abstände zwischen zwei Schwellen innerhalb eines Items, aber nicht über alle Items gleich groß																

20. Testkonstruktion **PPT** Warum ein Mess-Modell ? → Zählen ist nicht Messen !

Zuweisung von Zahlen zu Antwortmöglichkeiten ≠ Messen

<p>Messen → homomorphe Abbildung</p>	<p>„Zuordnen von Zahlen zu Eigenschafts- oder Fähigkeitsausprägungen, so dass die Relationen der Eigenschafts- oder Fähigkeitsausprägungen durch die Relationen der Zahlen abgebildet werden.“</p> <p style="text-align: right;"><i>(Markus Bühner, Einführung in die Testkonstruktion)</i></p>
---	---

Ziel einer psychologischen Messung wäre also ein Aussage wie z.B.

A ist doppelt so ängstlich wie B
A: Angst = 6
B: Angst = 3

Problem:
Zählen von Rohwerten führt selten zur genauen Bestimmung

Messmodell:
vom konkreten Zählen
zum abstrakten Messen

Beispiel

	A	B	
leicht	item 1	nein	ja
	item 2	nein	ja
	item 3	nein	ja
schwierig	item 4	ja	nein
	item 5	ja	nein
	item 6	ja	nein
gezählt:	3 Punkte	3 Punkte	

Die Punkte sind **nicht** gleichwert

Itemanalyse

im Rahmen der KTT

im Rahmen der PTT

<p>Itemschwierigkeit Trennschärfe</p>	<p>Itemparameter: Wie schwierig ist das Item ?</p> <p>↓</p> <p>Personenparameter: Welche Personenfähigkeit steht hinter der Lösung dieses Items ?</p> <p>+ andere Parameter</p> <p>↓</p> <p>Lösungswahrscheinlichkeit ?</p>
---	--

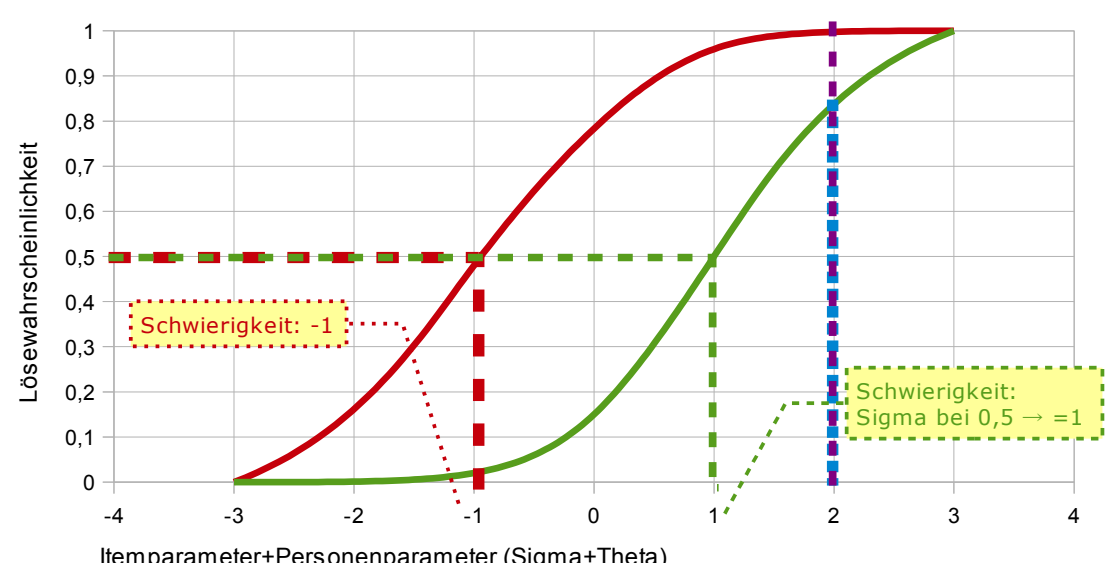
21. Testkonstruktion **PPT** Mess-Modelle (für dichotome Items)

Item Characteristic Curve



<p>Guttman</p>		<p><u>Lösungswahrscheinlichkeit: Werte 0 oder 1</u> → je nachdem, ob die Personenfähigkeit die Itemschwierigkeit übersteigt oder nicht</p> <p><u>Information:</u> nur Rangreihen → keine erschöpfende Statistik ohne Kenntnis des Antwortmusters</p>
<p>essenziell Tau-äquivalente Messung</p>		<p><u>Lösungswahrscheinlichkeit steigt linear an mit zunehmender Differenz zwischen Personenfähigkeit und Itemschwierigkeit</u></p> <p><u>Information:</u> Intervallskalenniveau der Summenwerte</p> <p>Problem: Werte unter 0 und über 1 sind möglich – können aber eigentlich nach Definition „Wahrscheinlichkeit“ nicht sein</p>
<p>logistische Regression</p>		<p><u>Lösungswahrscheinlichkeit steigt monoton an mit zunehmender Differenz zwischen Personenfähigkeit und Itemschwierigkeit</u></p> <p>Werte zwischen 0 und 1</p> <p><u>Information:</u> → zeigt auch für Personen mit geringen Fähigkeiten (entsprechend geringe, aber nicht keine) Lösungswahrscheinlichkeiten für schwierige Items an</p>

siehe Rasch-Modell

<p>Logit-Einheit</p>	<p>abstraktes, gemeinsames Maß für Item- und Personenparameter → ermöglicht additive Verknüpfung der Parameter zu „Lösewahrscheinlichkeit“</p>																															
	<p>$\theta - \sigma = \dots$</p>	<p>je höher Personenparameter Theta, desto größer die Chance, dass nach Abzug von Itemparameter Sigma ein „großer Wert“ (= hohe Lösewahrscheinlichkeit) bleibt</p>																														
<p>Methode</p> <table border="1" data-bbox="79 470 311 761"> <tr><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr> <tr><td>2</td><td>1</td><td>1</td><td>0</td><td>1</td></tr> <tr><td>3</td><td>1</td><td>0</td><td>0</td><td>1</td></tr> <tr><td>4</td><td>0</td><td>1</td><td>0</td><td>1</td></tr> <tr><td>5</td><td>0</td><td>1</td><td>0</td><td>0</td></tr> <tr><td>6</td><td>0</td><td>0</td><td>1</td><td>1</td></tr> </table>	1	2	3	4	5	2	1	1	0	1	3	1	0	0	1	4	0	1	0	1	5	0	1	0	0	6	0	0	1	1	<p>Schätzen der Itemparameter nach der cMLM <i>conditional Maximum-Likelihood-Methode</i> :</p> <p>Ausgangspunkt: beobachtete Datenmatrix</p> <p><i>Wie sehen wahrscheinlich die Modellparameter in der Population aus, die zu dieser Matrix geführt haben ?</i></p>	<p>Schätzen der Personenparameter mit Kenntnis der Itemparameter nach der gewichteten MLM</p> <p>→ kriteriumsorientierte Interpretation der Personenparameter auf Itemebene*</p> <p>→ stichprobenunabhängig</p> <p>*auf Testebene:</p> <ul style="list-style-type: none"> • zusammengefasste Items=Kriterium • mittlere Lösungswahrscheinlichkeit einer Person=Personenparameter
1	2	3	4	5																												
2	1	1	0	1																												
3	1	0	0	1																												
4	0	1	0	1																												
5	0	1	0	0																												
6	0	0	1	1																												
<p>Logit-Transformation</p> <p>Transformation der Lösewahrscheinlichkeit in eine abstrakte Maßeinheit</p> <p>(die dann über Personen-/Itemparameter aufgetragen wird)</p> <p>→</p> <table border="1" data-bbox="79 1478 311 1635"> <tr><td>$\theta = \sigma$</td><td>$p=0.5$</td></tr> <tr><td>$\theta > \sigma$</td><td>$p > 0,5$</td></tr> <tr><td>$\theta < \sigma$</td><td>$p < 0,5$</td></tr> </table>	$\theta = \sigma$	$p=0.5$	$\theta > \sigma$	$p > 0,5$	$\theta < \sigma$	$p < 0,5$	<p>1. Odds-Ratio /Wettquotient: $\frac{\text{Wahrscheinlichkeit, mit der eine Person das Item löst}}{\text{Wahrscheinlichkeit, mit der eine Person das Item nicht löst}}$</p> <p>2. Logarithmieren des Wettquotienten:</p> <ul style="list-style-type: none"> • Itemschwierigkeit: $\theta - \sigma = 0 \rightarrow \theta = \sigma \rightarrow$ Lösungswahrscheinlichkeit 50% • Wertebereich (theoretisch): -unendlich bis +unendlich → unendlich viele Items mit unendlich vielen Schwierigkeitsgraden → unendlich viele Personen mit unendlich vielen „Fähigkeitsgraden“ <p>meistens: Werte zwischen -3 und +3</p> <table border="1" data-bbox="510 1344 1516 1478"> <thead> <tr> <th>negative Werte</th> <th>positive Werte</th> </tr> </thead> <tbody> <tr> <td>leichte Items + Personen mit geringen Fähigkeiten</td> <td>schwierige Items + Personen mit hohen Fähigkeiten</td> </tr> </tbody> </table>  <p>Eine Person mit $\theta = 2$ löst das grüne Item ($\sigma = 1$) mit einer Wahrscheinlichkeit von 85%, das rote Item ($\sigma = -1$) mit einer Wahrscheinlichkeit von 100%</p>		negative Werte	positive Werte	leichte Items + Personen mit geringen Fähigkeiten	schwierige Items + Personen mit hohen Fähigkeiten																				
$\theta = \sigma$	$p=0.5$																															
$\theta > \sigma$	$p > 0,5$																															
$\theta < \sigma$	$p < 0,5$																															
negative Werte	positive Werte																															
leichte Items + Personen mit geringen Fähigkeiten	schwierige Items + Personen mit hohen Fähigkeiten																															

Ist das dichotome Rasch-Modell mit den Daten vereinbar ?

→ Auswahl von Items, die den Anforderungen des Messmodells entsprechen

Streudiagramm der geschätzten Itemparameter aus 2 Teilstichproben

z.B. unter- und überdurchschnittliche Leistungen im Intelligenztest durch Spaltung der Personenwerte aus der vorliegenden Stichprobe in zwei Teilstichproben am Median der Leistung

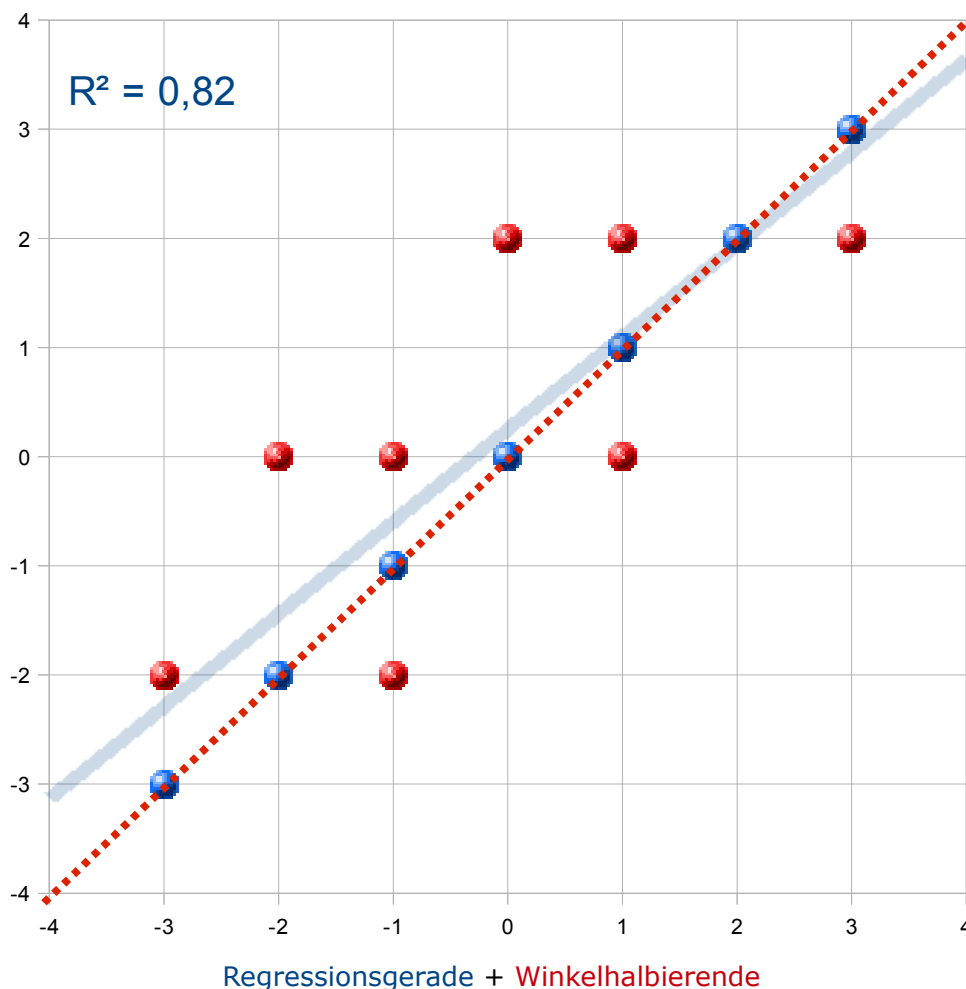
- Bestimmung der Itemparameter für die jeweiligen Teilstichproben
- Eintragen der Itemparameter in EIN Streudiagramm

↳ Bilden die Itemparameter eine Winkelhalbierende ?

↓ wenn ja

→ absolute Übereinstimmung zwischen den Itemparametern der beiden Teilstichproben

- Personenhomogenität
- Modellgeltung



hier weichen die roten Items alle mehr oder weniger stark von der Winkelhalbierenden ab... → müssten aus dem Test ausgeschlossen werden (bzw. verletzt der Gesamtest hier die Voraussetzungen des Rasch-Modells)

Nachteile

- zahlreiche Möglichkeiten zur Aufspaltung der Stichprobe → unterschiedliche Ergebnisse für andere Aufteilungen möglich
- Abweichung von der Winkelhalbierenden aufgrund von Messfehlern möglich
- rein deskriptives Verfahren kein globaler Signifikanztest

Vorteile

- hohe Anschaulichkeit.
- ermöglicht Entdeckung auffälliger Items
- Abweichungen können interpretiert werden: wie werden Items von unterschiedlichen Teilstichproben „erlebt“ ?

<p>Signifikanztests</p> <p>strengere Voraussetzungen und daher kritisch in der Anwendung</p>	<p>Andersen-Likelihood-Quotienten-Test</p>	<p>CML-Schätzungen für beide Teilstichproben → Signifikanztest auf Unterschiedlichkeit → bei Beibehalten der H_0 „Gleiche Itemparameter für alle Teilpopulationen“ = Modellkonformität</p>
	<p>Pearson-χ^2-Test</p> <p>Bootstrap-Methode</p>	<p>Vergleichen von beobachteten und erwarteten Häufigkeiten</p> <p>Test auf Basis einer wiederholten Ziehung und statistischer Auswertung von Unterstichproben aus einer Ausgangsstichprobe</p>
<p>Modellvergleiche</p>	<p>Welches (Mess-)Modell passt am besten zu den empirischen Daten ?</p>	

→
bei

Bestätigung des dichotomen Rasch-Modells

(→ Modelltest wird nicht signifikant)

gelten folgende Annahmen:

1. lokale stochastische Unabhängigkeit
2. erschöpfende Statistik
3. Itemhomogenität mit gleicher Trennschärfe = 1
4. spezifische Objektivität

1. **lokale stochastische Unabhängigkeit**
 Prüfung:
 Bei einer bestimmten (konstantgehaltenen) Merkmalsausprägung weisen die Items keine Korrelation auf.
 (vgl KTT: unkorrelierte Messfehler)
 →
 Leistung bei einem Item ist komplett auf die Fähigkeitsausprägung der latenten Variablen zurückzuführen
 →
 Wahrscheinlichkeiten für alle Items können multipliziert werden

Achtung:
 Lokale stochastische Unabhängigkeit \neq korrelative Unabhängigkeit
 → Items können miteinander korrelieren (!)
wär ja in Sachen Itemhomogenität auch doof, wenn sie das nicht dürften...
- Lokale stochastische Unabhängigkeit ist eine **Voraussetzung, aber kein Beweis** von Eindimensionalität!
2. **Summenwert der Itemantworten**
 als **erschöpfende Statistik** der Personenfähigkeit:
 →
 Summenwert einer Person liefert alle Informationen über die Fähigkeitsausprägung
3. **Itemhomogenität:**
 homogene Items unterscheiden sich nur in der Schwierigkeit,
Trennschärpen sind gleich (= 1) → parallele Itemkurven, an x-Achse verschoben
4. **spezifische Objektivität**
 - Vergleiche zwischen Personen sind spezifisch objektiv (auf die verwendeten Items bezogen)

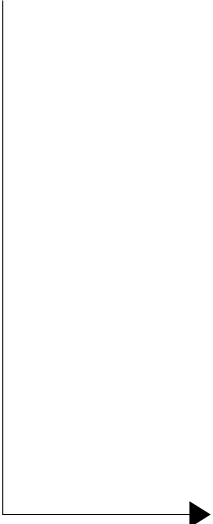
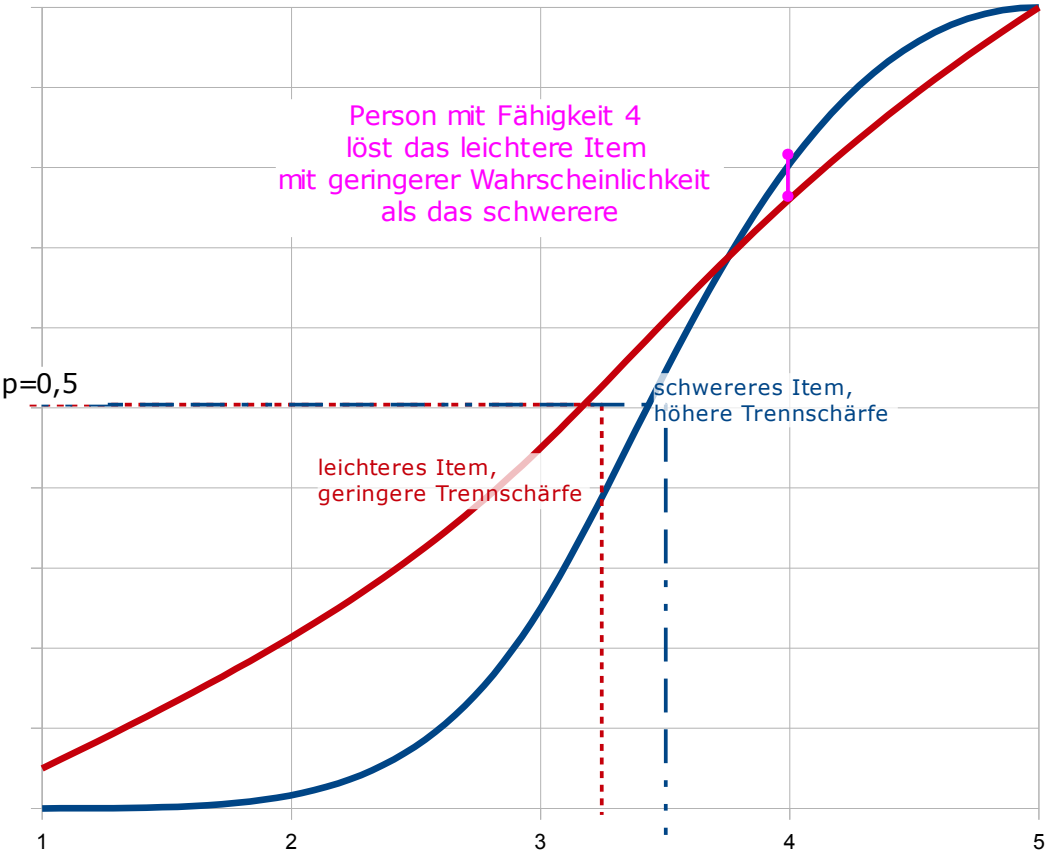
 Die Fähigkeit einer Person ist nur davon abhängig, wie viele Items sie gelöst hat – nicht welche Items.
 - Vergleiche zwischen Items sind spezifisch objektiv (auf die „verwendeten“ Personen bezogen)

 Die Schwierigkeit eines Items ist nur davon abhängig, von wie vielen Personen es gelöst wurde – nicht von welchen Personen.

→ ermöglicht adaptives Testen

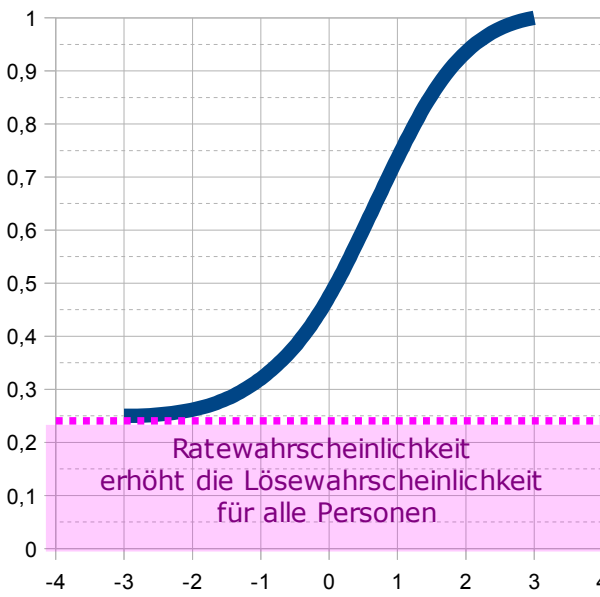
25. Testkonstruktion PPT Mess-Modelle: 2PL-Modell (Birnbaum-Modell)

Wenn das dichotome Rasch-Modell wegen mangelnder Itemhomogenität nicht zu den Daten passt, passt vielleicht wenigstens das hier...

<p>Itemparameter + Personenparameter +</p> <p>Trennschärfe-/ Diskriminationsparameter</p>	<p>Trennschärfe = Tangente der ICC → Steigung der ICC am Wendepunkt ($p=0,5$)</p> <p>→ je höher die Trennschärfe, desto steiler der Anstieg</p> <p>Vgl. Trennschärfe KTT = Korrelation Item / Skala</p> <p>Werte: 0 bis +unendlich</p>
<p>Problem bei unterschiedlichen Trennschärfen</p> 	<p>leichtere Items mit höherer Trennschärfe werden u.U. von Personen mit eigentlich höheren Fähigkeiten mit geringerer Wahrscheinlichkeit gelöst als schwierigere Items mit geringerer Trennschärfe</p> <p>→ unterschiedliche Trennschärfen müssten also eigentlich in die Itemschwierigkeit mit eingerechnet werden</p> <p>→ Summenwert ≠ erschöpfende Statistik: wenn ich die Trennschärfe nicht kenne, kann ich allein auf Grundlage der Anzahl der gelösten Items keine Aussage zur Personenfähigkeit machen</p> <p>vgl. dichotomes Rasch-Modell: Trennschärfe = 1 → Summenwert der Items als erschöpfende Statistik der Personenfähigkeit</p> 

26. Testkonstruktion PPT Mess-Modelle: 3PL-Modell (Rate-Birnbaum-Modell)

Wenn das 2PL-Modell wegen zusätzlicher Ratewahrscheinlichkeit nicht zu den Daten passt, passt vielleicht wenigstens das hier...

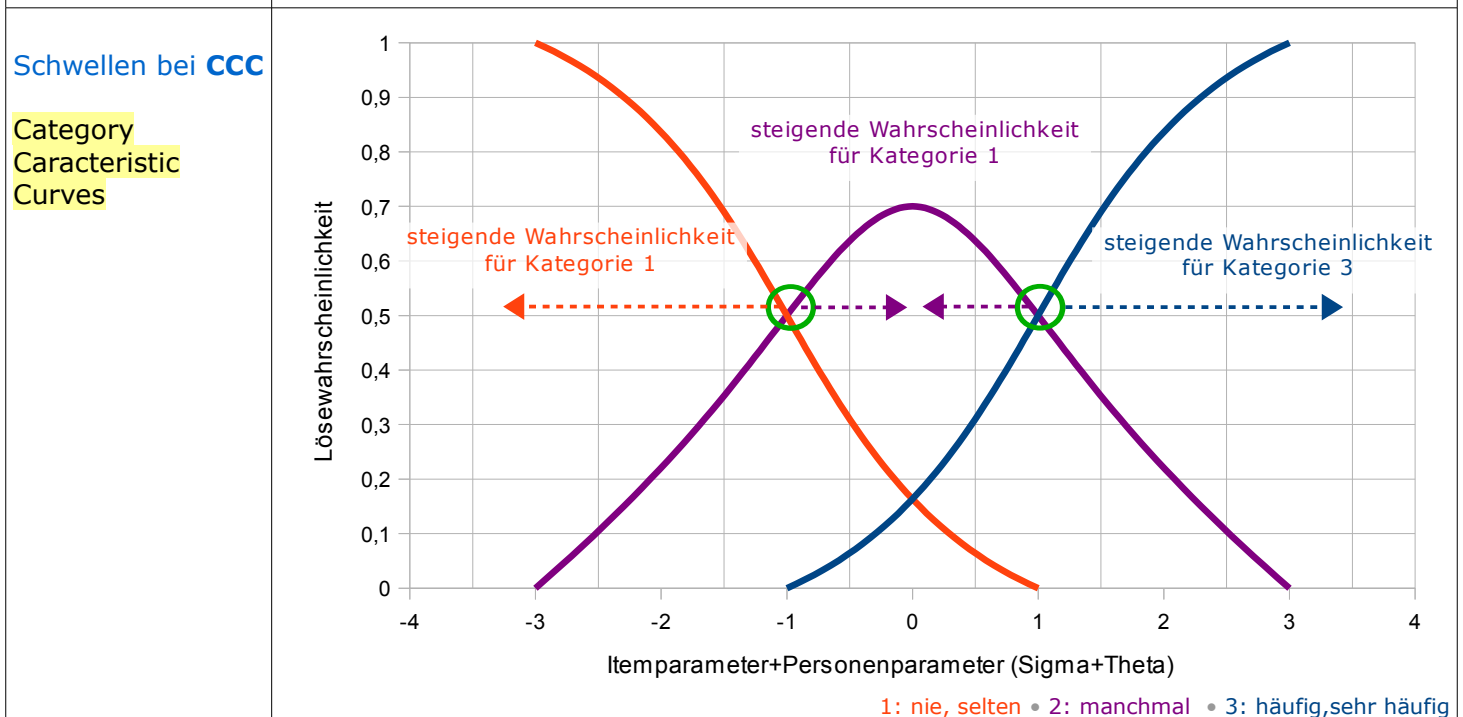
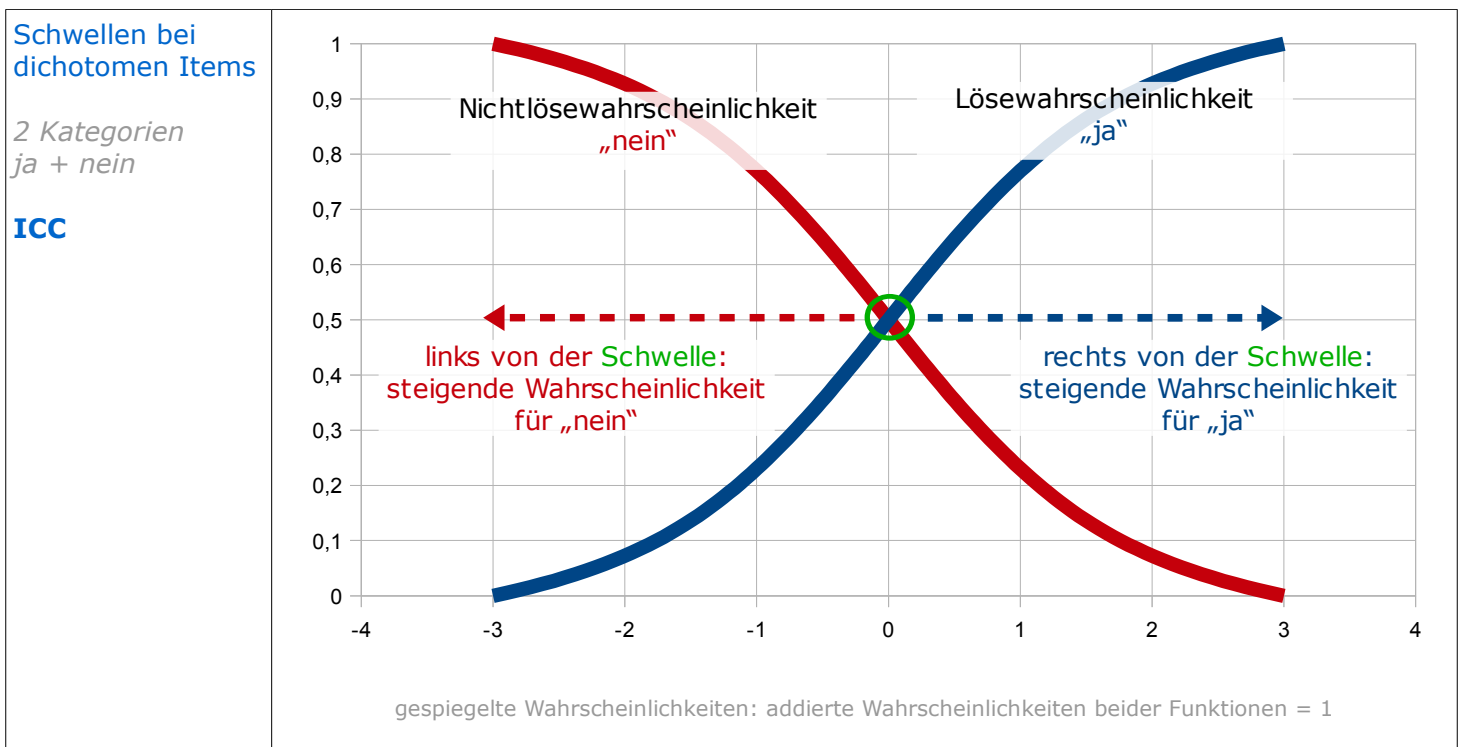
<p>Itemparameter + Personenparameter + Trennschärfe-/ Diskriminationsparameter +</p> <p>Rateparameter</p>	<p>Ratewahrscheinlichkeit:</p> <p>Bsp: Item mit 4 Antwortalternativen → richtig: 1 aus 4 → Ratewahrscheinlichkeit: 25%</p>	
<p>Problem</p>	<p>Ratewahrscheinlichkeit hängt bei jedem Item nicht nur von der Anzahl, sondern auch von der Qualität der Distraktoren ab</p> <p>→ Rateparameter ist nicht über alle Items konstant</p>	
<p>→ besser</p>	<p>... als nachträgliche Einführung eines Rateparameters:</p> <p>Ratewahrscheinlichkeit schon bei Konstruktion des Testes durch geeignete Maßnahmen minimieren, um nicht auf dieses Messmodell zurückgreifen zu müssen !</p>	

27. Testkonstruktion **PPT** Mess-Modelle: **ordinales Rasch-Modell** → Schwellenkonzept

<p>Voraussetzung</p>	<p>geordnete Antwortkategorien → überschneidungsfreie Schwellen</p> <p style="text-align: right;">nie • selten • manchmal • häufig • sehr häufig</p>
<p>Annahme</p>	<p>mit steigender Merkmalsausprägung steigt die Wahrscheinlichkeit, eine höhere Antwortkategorie zu wählen</p> <p>PTT-Sprache: je höher der Personenparameter im Vergleich zur Antwortkategorie, -desto wahrscheinlicher wird diese Kategorie gewählt</p>

→ **Schwellenkonzept:**

Schwellenparameter = gleiche Antwortwahrscheinlichkeiten auf beiden Seiten des Kurvenschnittpunktes



28. Testkonstruktion PPT Mess-Modelle: Mixed-Rasch-Modell

Bei homogenen Items könnte das Problem vielleicht an der Heterogenität der Personen liegen...

<p>quantifizieren + klassifizieren</p>	<p>→ Messen der Merkmalsausprägung + → Einordnen von Probanden in Klassen (z.B. nach unterschiedlichen Lösungsstrategien)</p>																																		
<p>unterschiedliche Antwortmuster als Hinweise auf fehlende Eindimensionalität</p>	<p>→ Ermitteln von Personengruppen, die sich in ihren Antwortmustern maximal unterscheiden</p> <p>→ Verdacht: diese Gruppen unterscheiden sich nicht (nur) in dem untersuchten Merkmal</p> <p>→ in jeder Klasse wird dabei eine andere Fähigkeit erfasst: die Personen in den unterschiedlichen Klassen nutzen nicht dieselben Eigenschaften oder Lösungsstrategien bei der Bearbeitung des Tests</p> <p>→ zusätzliche Dimension → „Latent Trait“ → “Latent-Trait-Modell“</p> <p>Beispiel:</p> <table border="1" data-bbox="395 949 1528 1281"> <thead> <tr> <th></th> <th>Item 1</th> <th>Kategorie 1</th> <th>Kategorie 2</th> <th>Kategorie 3</th> <th></th> </tr> </thead> <tbody> <tr> <td rowspan="3">Extremkreuzer</td> <td>Person A</td> <td>1</td> <td>0</td> <td>0</td> <td rowspan="3">beantworten Item aufgrund von Variable X</td> </tr> <tr> <td>Person B</td> <td>0</td> <td>0</td> <td>1</td> </tr> <tr> <td>Person C</td> <td>1</td> <td>0</td> <td>0</td> </tr> <tr> <td rowspan="3">Mittelkreuzer</td> <td>Person D</td> <td>0</td> <td>1</td> <td>0</td> <td rowspan="3">beantworteten Item aufgrund von Variable Y</td> </tr> <tr> <td>Person E</td> <td>0</td> <td>1</td> <td>0</td> </tr> <tr> <td>Person F</td> <td>0</td> <td>1</td> <td>0</td> </tr> </tbody> </table>		Item 1	Kategorie 1	Kategorie 2	Kategorie 3		Extremkreuzer	Person A	1	0	0	beantworten Item aufgrund von Variable X	Person B	0	0	1	Person C	1	0	0	Mittelkreuzer	Person D	0	1	0	beantworteten Item aufgrund von Variable Y	Person E	0	1	0	Person F	0	1	0
	Item 1	Kategorie 1	Kategorie 2	Kategorie 3																															
Extremkreuzer	Person A	1	0	0	beantworten Item aufgrund von Variable X																														
	Person B	0	0	1																															
	Person C	1	0	0																															
Mittelkreuzer	Person D	0	1	0	beantworteten Item aufgrund von Variable Y																														
	Person E	0	1	0																															
	Person F	0	1	0																															
<p>aber: Gültigkeit des Rasch-Modells innerhalb einer Klasse</p>	<p>→ innerhalb jeder Klasse wird eindimensional eine Fähigkeit gemessen</p>																																		