Grundlagen der Testtheorie
WS 2020/21

5. Item Response Theorie
07.12.2020

Prof. Dr. Eunike Wetzel

Semesterplan

Sitzung	Termin	Thema
1	02.11.	Grundlagen & Gütekriterien
2	09.11.	Schritte der Testkonstruktion: Übersicht Konstruktdefinition & Itemgenerierung
3	16.11.	Erstellung eines Testentwurfs
4	23.11.	Klassische Testtheorie
5	**07.12.**	**Item Response Theorie**
6	14.12.	Exploratorische Faktorenanalyse 1
7	04.01.	Exploratorische Faktorenanalyse 2

Item Response Theorie (IRT)

- Modelle im Rahmen der IRT modellieren den Zusammenhang zwischen den Antworten einer Person auf die Items und der zugrundeliegenden latenten Variable
- Ziel: Fundierte Aussagen über die Ausprägung der Personen auf der latenten Variable treffen
- Im Mittelpunkt der Betrachtung stehen die Antworten der Personen auf die Items → Modellierung auf Itemebene
- In die Schätzung der Ausprägung von Personen auf der latenten Variable gehen sowohl die Antworten der Personen als auch die Eigenschaften der Items ein

Warum IRT?

- LATENTE Modellierung
  - Beobachtungen werden nicht mit latenten Konstrukten gleichgesetzt
  - Systematische Messfehlereinflüsse (z. B. Antwortstile) können explizit modelliert werden
- IRT-Modelle sind falsifizierbar (Modellgeltung ist überprüfbar)
- Abhängigkeit der Messpräzision vom Traitlevel wird berücksichtigt
- Voraussetzungen für die Verwendung von Summenscores oft nicht erfüllt

Anwendungsbereiche der IRT

- Konstruktion neuer Tests (z. B. Trierer Integriertes Persönlichkeitsinventar; Becker, 2003)
- Überprüfung etablierter Tests (z. B. Untersuchungen zu NEO-PI-R, IST-2000 R etc.)
- Large Scale Assessments in der empirischen Bildungsforschung (z. B. PISA)
- Erstellung paralleler Testformen
- Adaptives Testen (z. B. Adaptives Intelligenz Diagnostikum, Kubinger & Holocher-Ertl, 2014)
- …

Übersicht IRT-Modelle

IRT-Modelle lassen sich unterteilen in

1. Modelle mit einer quantitativen Personenvariable
   - Rasch Modell, 1-parametrisches logistisches (PL) Modell, 2PL, 3PL, 4PL Modelle
   - Jeweils für dichotome Daten (z.B. Rasch Modell) und polytome Daten (z.B. Partial Credit Modell)
2. Modelle mit einer qualitativen Personenvariable: Latente Klassenanalyse
3. Modelle mit beidem: Mixed Rasch Modelle
4. Modelle mit mehr als einer quantitativen Personenvariable: Mehrdimensionale Modelle

Item Response Theorie

1. **Hinführung zum Rasch Modell**
    1. Summenscore
    2. Lösungswahrscheinlichkeit
2. Rasch Modell
    1. Modellgleichung
    2. Item Characteristic Curve
    3. Annahmen und Eigenschaften
    4. Normierung und Messniveau
    5. Maximum Likelihood Schätzung
3. Beispiel Algebratest

Rasch Modell

- Das Rasch Modell (RM) geht von einer kontinuierlichen latenten Variable aus, die das interessierende Konstrukt repräsentiert
- Die latente Variable wird mit dichotomen Items erfasst

Example: $\sqrt{9+16} = ?$

$\sqrt{9+16} = 5$ (1, correct)
$\sqrt{9+16} = 25$ (0, wrong)

Example: "I don't talk a lot"

☐ *agree* (1)
☐ *disagree* (0)

Datenmatrix

Datenmatrix für einen Fähigkeitstest mit 6 Items

		Items					
		1	**2**	**3**	**4**	**5**	**6**
Personen	**1**	0	1	0	1	0	1
	2	0	1	1	0	1	1
	3	0	1	1	1	0	0
	4	1	0	0	1	0	0

Datenmatrix

Allgemeine Datenmatrix:

		Items						
		1	**2**	**…**	**j**	**…**	**m-1**	**m**
Personen	**1**	$u_{1,1}$	$u_{1,2}$	…	$u_{1,j}$	…	$u_{1,m-1}$	$u_{1,m}$
	2	$u_{2,1}$	$u_{2,2}$	…	$u_{2,j}$	…	$u_{2,m-1}$	$u_{2,m}$
	:	…	…	…	…	…	…	…
	i	$u_{i,1}$	$u_{i,2}$	…	$u_{i,j}$	…	$u_{i,m-1}$	$u_{i,m}$
	:	…	…	…	…	…	…	…
	n-1	$u_{n-1,1}$	$u_{n-1,2}$	…	$u_{n-1,j}$	…	$u_{n-1,m-1}$	$u_{n-1,m}$
	n	$u_{n,1}$	$u_{n,2}$	…	$u_{n,j}$	…	$u_{n,m-1}$	$u_{n,m}$

$u_{i,j}$: Antwort von Person *i* auf Item *j*

Es gibt *n* Zeilen (Personen) und *m* Spalten (Items).

1.1 Summenscore

$$\sum_{j=1}^{m} u_{ij} = r_i$$

- Auch „Personenscore“
- Wird als Fähigkeit bzw. Traitausprägung der Person interpretiert
- Ob Summe der Itemantworten ein adäquates Maß für die Fähigkeit oder Traitausprägung der Person darstellt, ist eine zentrale Frage psychologischer Testtheorie
- Alternative: Antwortmuster betrachten

Scoreverteilung

r	0	1	2	3	4	5
n_r	2	2	3	4	0	1

- r = Score; n_r = Häufigkeit des Scores
- Häufigkeitsverteilung der Summenscores
- Alle Personen, die dieselbe Anzahl von Items gelöst haben, werden als gleich fähig betrachtet
- Lässt Verteilungsform der Stichprobe erkennen (gleichmäßig, Deckeneffekt, Bodeneffekt)

1.2 Lösungswahrscheinlichkeit

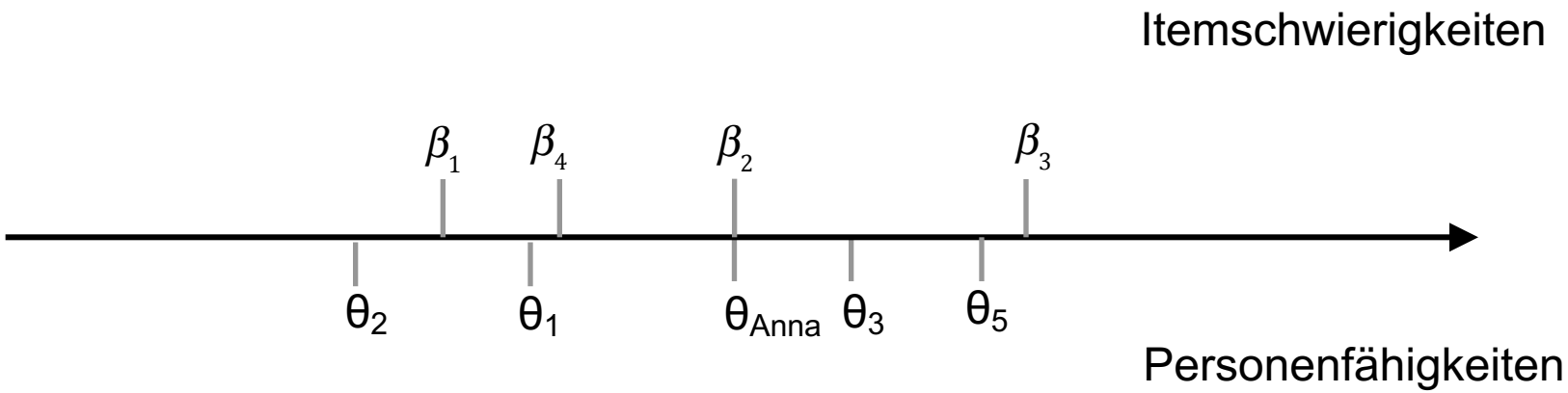
- Ergebnis der Person im Test hängt nicht deterministisch von ihrer Fähigkeit ab, sondern es ist auch immer etwas Zufall im Spiel
- In der IRT wird deshalb die Lösungswahrscheinlichkeit betrachtet: Die Wahrscheinlichkeit, mit der eine Person ein Item richtig löst
- Neben den beobachteten Itemantworten u_{ij} gibt es noch die Zufallsvariable U_{ij}
- Die Wahrscheinlichkeit (Probability = P), dass Person *i* bei der Bearbeitung von Item *j* das Ergebnis u_{ij} erzielt, wird bezeichnet als

$$P(U_{ij} = u_{ij})$$

1.2 Lösungswahrscheinlichkeit

- $P(U_{ij} = u_{ij})$ ist abhängig von der Fähigkeit der Person *i* und der Schwierigkeit des Items *j*
- θ: latente Variable (Trait oder Fähigkeit, allgemein: Personenparameter)
- β: Itemschwierigkeit
- Personenfähigkeit und Itemschwierigkeit werden auf einer gemeinsamen Skala geschätzt, deshalb können sie direkt zueinander in Beziehung gesetzt werden

Skala der Item- und Personenparameter



Je größer θ verglichen mit β, um so wahrscheinlicher wird eine korrekte Antwort

Je kleiner θ verglichen mit β, um so wahrscheinlicher wird eine falsche Antwort

Item Response Theorie

1. Hinführung zum Rasch Modell
    1. Summenscore
    2. Lösungswahrscheinlichkeit
2. **Rasch Modell**
    1. Modellgleichung
    2. Item Characteristic Curve
    3. Annahmen und Eigenschaften
    4. Normierung und Messniveau
    5. Maximum Likelihood Schätzung
3. Beispiel Algebratest

2.1 Rasch Modell

- Modellgleichung:

$$P(u_{ij} = 1 \mid \theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}$$

- Eigenschaften:
  - Lösungswahrscheinlichkeit hängt von der Fähigkeit der Person θ_i ab
  - Lösungswahrscheinlichkeit hängt von der Schwierigkeit des Items β_j ab
  - Funktion steigt mit der Personenfähigkeit an
  - Ergebnis ist eine Wahrscheinlichkeit, d.h. die Grenzen liegen zwischen 0 und 1

2.1 Beispiel Rasch Modell

Beispiel: $\theta_i = 3$ und $\beta_j = 2$

$$P(u_{ij} = 1 \mid \theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}$$

$$P(u_{ij} = 1 \mid \theta_i = 3, \beta_j = 2) = \frac{e^{3-2}}{1 + e^{3-2}} = \frac{e^1}{1 + e^1} \approx 0.73$$

Beispiel: $\theta_i = 4$ und $\beta_j = 2$

$$P(u_{ij} = 1 \mid \theta_i = 4, \beta_j = 2) = \frac{e^{4-2}}{1 + e^{4-2}} = \frac{e^2}{1 + e^2} \approx 0.88$$

2.1 Beispiel Rasch Modell

Beispiel: $\theta_i = 2$ und $\beta_j = 3$

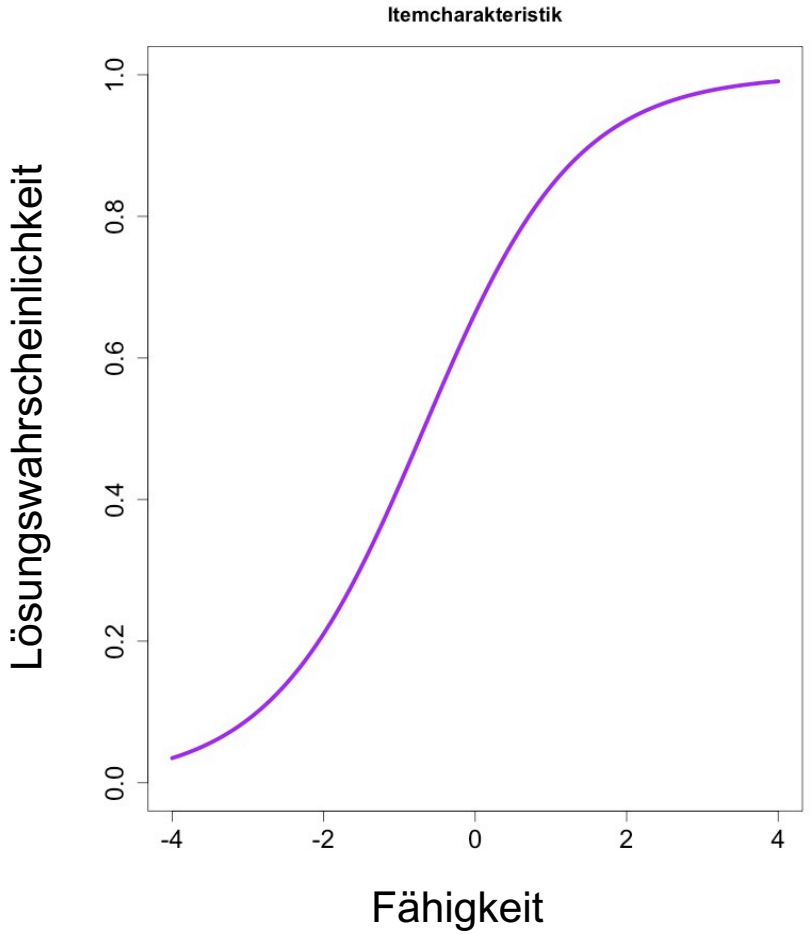
$$P(u_{ij} = 1 \mid \theta_i, \beta_j) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}$$

$$P(u_{ij} = 1 \mid \theta_i = 2, \beta_j = 3) = \frac{e^{2-3}}{1 + e^{2-3}} = \frac{e^{-1}}{1 + e^{-1}} \approx 0.27$$

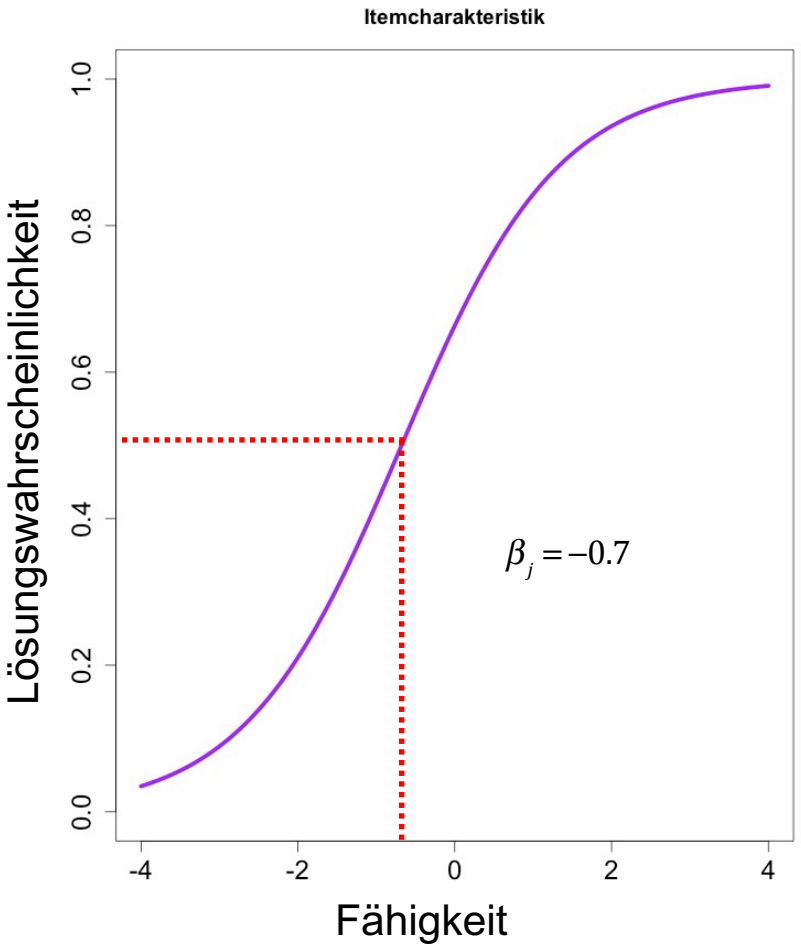
2.2 Logistische Funktion

Der Bruch $\frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}$ beschreibt eine logistische Funktion

Die logistische Funktion stellt den Verlauf der Lösungswahrscheinlichkeit für ein Item in Abhängigkeit von der Personenfähigkeit dar
→ Item Characteristic Curve (ICC)



2.2 Item Characteristic Curve (ICC)



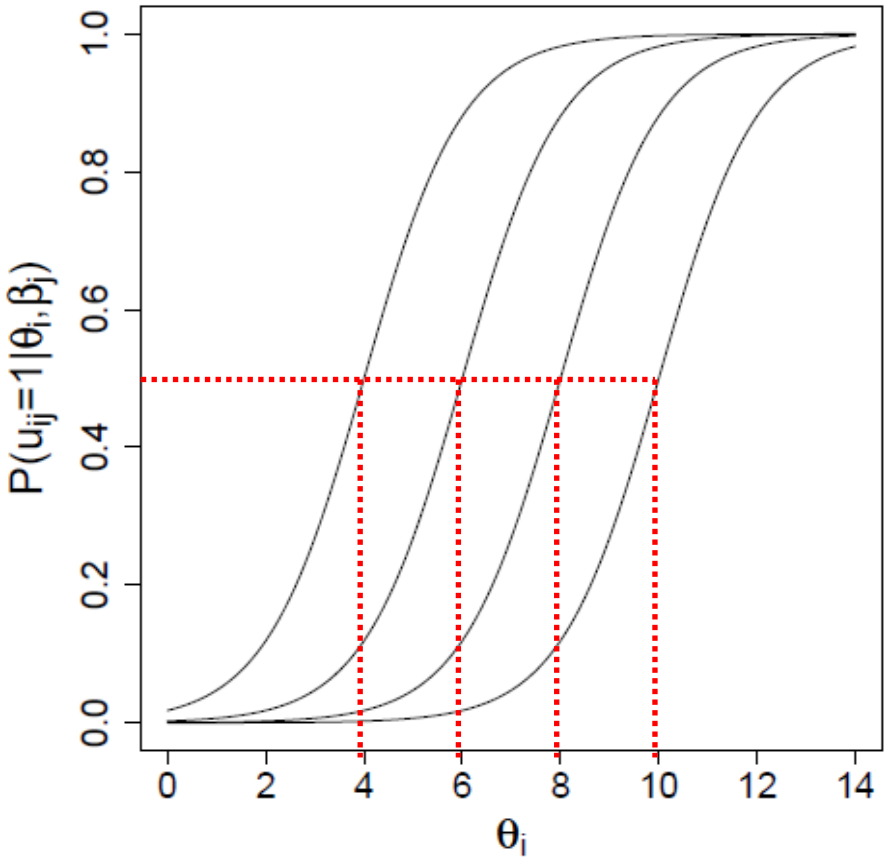
$$\textbf{Ist } \theta_i - \beta_j > 0, \textbf{ wird } P(u_{ij} = 1 \mid \theta_i, \beta_j) > 0.5$$

$$\textbf{Ist } \theta_i - \beta_j < 0, \textbf{ wird } P(u_{ij} = 1 \mid \theta_i, \beta_j) < 0.5$$

$$\textbf{Ist } \theta_i - \beta_j = 0, \textbf{ wird } P(u_{ij} = 1 \mid \theta_i, \beta_j) = 0.5$$

Die Itemschwierigkeit ist definiert als der Punkt auf dem Kontinuum der Fähigkeit, an dem die Lösungswahrscheinlichkeit 0.5 beträgt.

2.2 ICCs für Items mit unterschiedlichen Schwierigkeiten



Die ICCs verlaufen parallel zueinander, da es im Rasch Modell nur einen Parameter für die Items gibt (die Schwierigkeit)

Abb. 2.3 aus Strobl (2012)

2.1 Rasch Modell

- Gegenwahrscheinlichkeit (Wahrscheinlichkeit einer falschen Lösung):

$$P(u_{ij} = 0 \mid \theta_i, \beta_j) = \frac{1}{1 + e^{\theta_i - \beta_j}}$$

- Lösungswahrscheinlichkeit und Gegenwahrscheinlichkeit addieren sich zu 1:

$$P(u_{ij} = 1 \mid \theta_i, \beta_j) + P(u_{ij} = 0 \mid \theta_i, \beta_j) = 1$$

- Gemeinsame Formel für beides:

$$P(U_{ij} = u_{ij} \mid \theta_i, \beta_j) = \frac{e^{u_{ij} \cdot (\theta_i - \beta_j)}}{1 + e^{\theta_i - \beta_j}}$$

2.3 Annahmen und Eigenschaften des Rasch Modells

1. Suffiziente Statistiken
2. Eindimensionalität
3. Lokale stochastische Unabhängigkeit
4. Spezifische Objektivität

2.3.1 Suffiziente Statistiken

- Suffizient = erschöpfend
- Suffiziente Statistiken schöpfen die ganze in den Originaldaten vorhandene Information aus
- Ist eine Statistik suffizient, ist die Datenaggregation legitim und nicht mit einem Verlust an diagnostischer Information verbunden

2.3.1 Suffiziente Statistiken

		Items						
		1	**2**	**3**	**4**	**5**	**6**	r_i
Personen	**1**	0	1	0	1	0	1	3
	2	0	1	1	0	1	1	4
	3	0	1	1	1	0	0	3
	4	1	0	0	1	0	0	2
	s_j	1	3	2	3	1	2	

Im Rasch Modell ist die Zeilensumme r_i eine suffiziente Statistik für den Personenparameter θ_i und die Spaltensumme s_j eine suffiziente Statistik für den Itemparameter β_j

Rückblick: Summenscore

- Gilt das Rasch Modell, ist der Summenscore (Zeilensumme r_i) eine suffiziente Statistik zur Beschreibung der Person
- Die Betrachtung des Antwortmusters würde keine Informationen liefern, die über den Summenscore hinausgehen
- Mit einer Analyse zur Geltung des Rasch Modells kann überprüft werden, ob es gerechtfertigt ist, Summenscores zu verwenden
- In der KTT wird lediglich angenommen, dass Summenscores suffiziente Statistiken seien, in der IRT kann diese Annahme getestet werden!

2.3.2 Eindimensionalität

- Die Items im Test erfassen *ein* latentes Konstrukt
- Antworten der Personen auf die Items werden nur durch ein Konstrukt bestimmt
- Perfekte Eindimensionalität ist sehr schwer zu erreichen
- Häufig spricht man auch von „essenzieller Eindimensionalität“ wenn es einen dominanten Faktor gibt, der die Itemantworten bestimmt

2.3.3 Lokale stochastische Unabhängigkeit

- Lokale Unabhängigkeit in der KTT: Itemantworten sind unter Kontrolle der Traitausprägung unabhängig voneinander
- In der IRT bezieht sich die Forderung der lokalen Unabhängigkeit auf die Lösungswahrscheinlichkeiten der Items: Lokale stochastische Unabhängigkeit
- Lokale stochastische Unabhängigkeit ist gegeben, wenn gilt:

$$P(U_1 = u_1, \ldots, U_m = u_m \mid \theta_i) = \prod_{j=1}^{m} P(U_j = u_j \mid \theta_i)$$

2.3.3 Lokale stochastische Unabhängigkeit

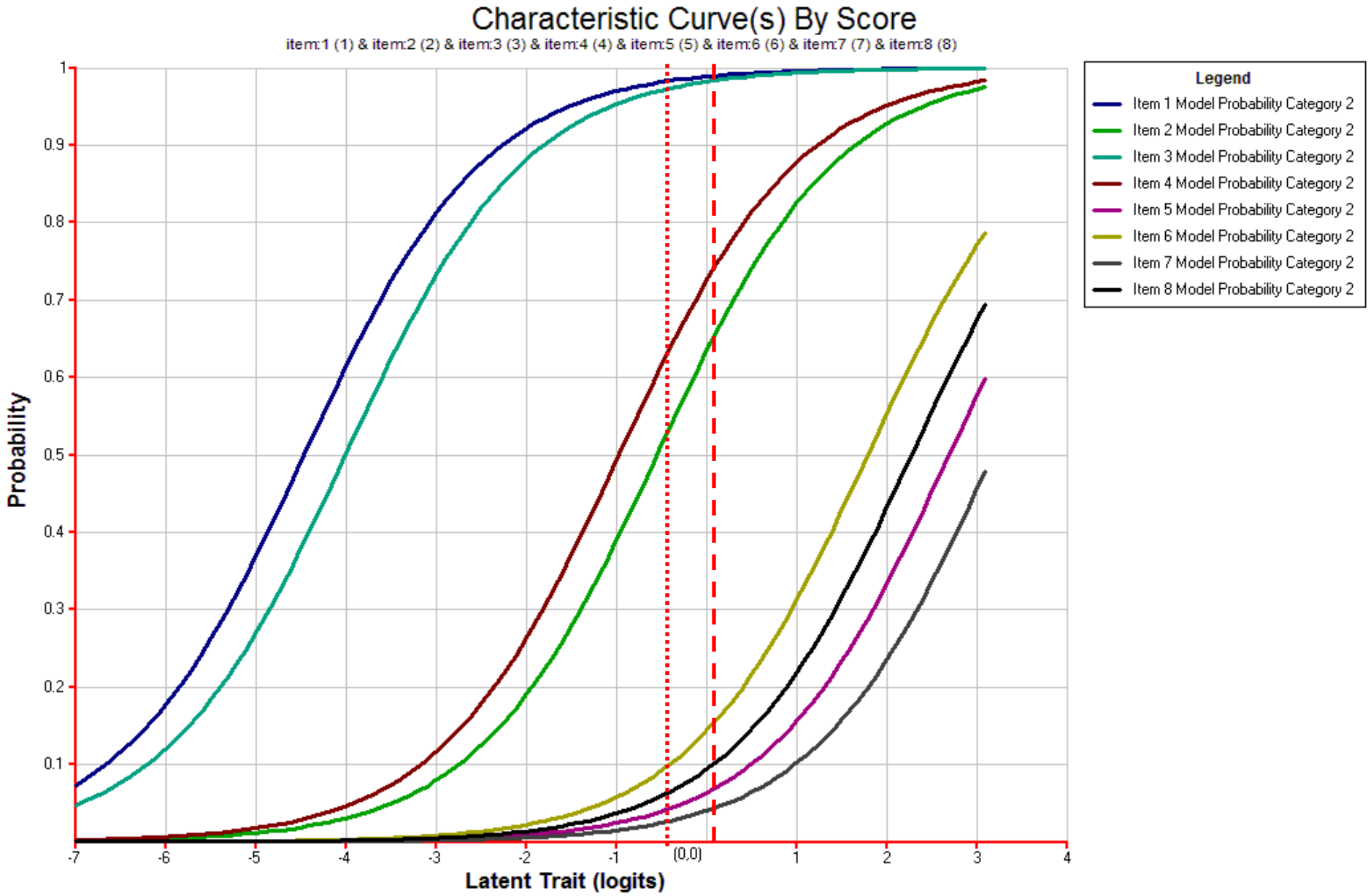
- Die Wahrscheinlichkeit, mit der eine Person ein Item richtig löst, muss unabhängig von den Lösungswahrscheinlichkeiten dieser Person für die anderen Items sein und unabhängig von den Lösungswahrscheinlichkeiten anderer Personen für das Item sein
- Lösungswahrscheinlichkeit für ein Item darf sich nicht durch das Lösen/Nicht-Lösen eines anderen Items verändern
- Lokal: Unabhängigkeit gilt nur, solange man eine Person (oder mehrere Personen mit der gleichen Fähigkeit) betrachtet

2.3.3 Lokale stochastische Unabhängigkeit

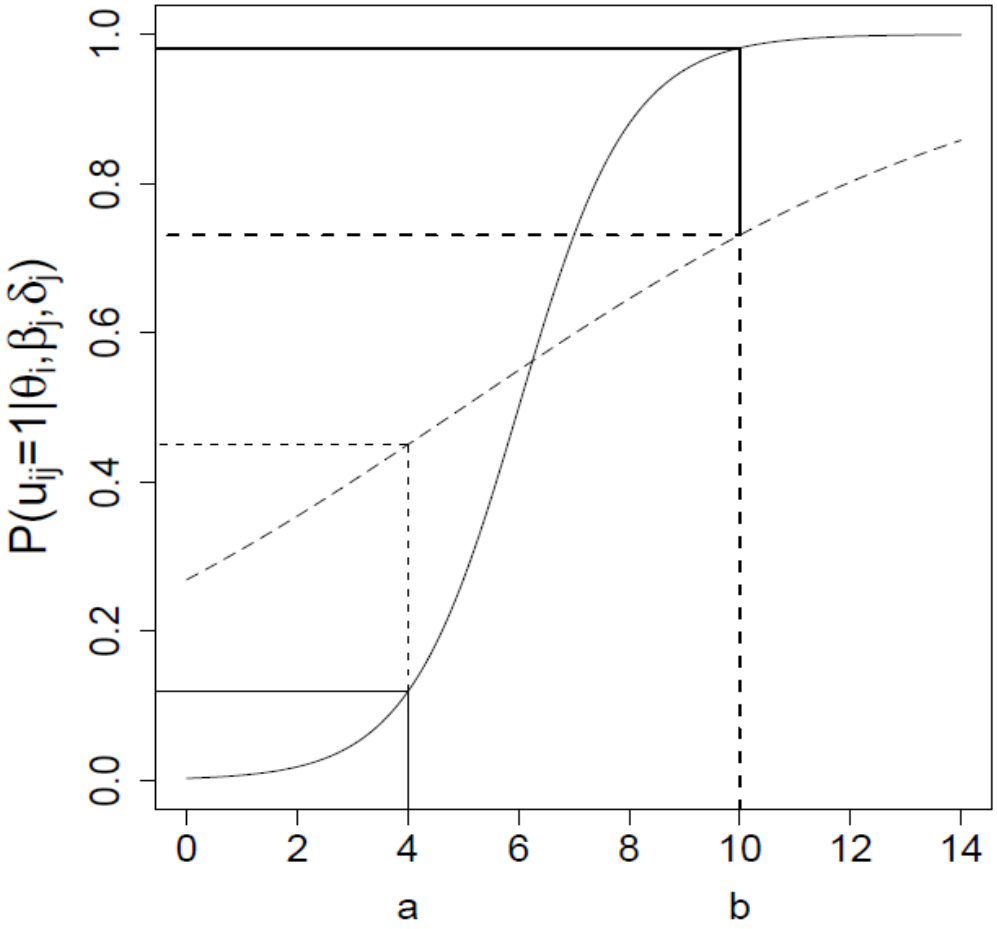
- Verletzungen der lokalen stochastischen Unabhängigkeit treten auf, wenn
  - Lerneffekte bei der Testung auftreten
  - Das den Itemantworten zugrundeliegende Konstrukt mehrdimensional ist
  - Konstruktfremde Variablen ebenfalls einen Einfluss auf die Itemantworten ausüben (z.B. Antwortstile)

2.3.4 Spezifische Objektivität

- Grundidee: Vergleiche zwischen Personen sollten nicht von den spezifischen vorgegebenen Items abhängen und Vergleiche zwischen Items sollten nicht von den spezifischen getesteten Personen abhängen
- Im Rasch Modell:
  - *Iteminvariante Personenparameter*: Unabhängig von den betrachteten Items sind die Personen nach ihren Fähigkeiten immer gleich geordnet
  - *Stichprobeninvariante Itemparameter*: Unabhängig von den betrachteten Personen sind die Items nach ihren Schwierigkeiten immer gleich geordnet


Characteristic Curve(s) By Score
item:1 (1) & item:2 (2) & item:3 (3) & item:4 (4) & item:5 (5) & item:6 (6) & item:7 (7) & item:8 (8)
Probability
Latent Trait (logits)
(0,0)
Legend
Item 1 Model Probability Category 2
Item 2 Model Probability Category 2
Item 3 Model Probability Category 2
Item 4 Model Probability Category 2
Item 5 Model Probability Category 2
Item 6 Model Probability Category 2
Item 7 Model Probability Category 2
Item 8 Model Probability Category 2

2.3.4 Spezifische Objektivität



- Spezifische Objektivität wird im Rasch Modell durch die parallelen ICCs gewährleistet
- Bei Modellen, in denen die ICCs unterschiedliche Steigungen haben können (Unterschiede in der Trennschärfe), ist keine spezifische Objektivität gegeben

Abb. 2.6 aus Strobl (2012)

Überprüfung der Annahmen

- Globale Modellgeltungstests prüfen, ob das Rasch Modell als Ganzes (mit allen Annahmen) auf die Daten passt
- Mithilfe von Itemfit-Statistiken kann auch beurteilt werden, wie gut einzelne Items zum Modell passen
- Diese Statistiken können in der Testkonstruktion zur Itemselektion eingesetzt werden

Item Response Theorie

1. Hinführung zum Rasch Modell
    1. Summenscore
    2. Lösungswahrscheinlichkeit
2. Rasch Modell
    1. Modellgleichung
    2. Item Characteristic Curve
    3. Annahmen und Eigenschaften
    4. **Normierung & Messniveau**
    5. Maximum Likelihood Schätzung
3. Beispiel Algebratest

2.4 Normierung & Messniveau

Warum Normierung?

- Die Skala der Personenfähigkeiten und Itemschwierigkeiten ist eine latente Skala, sie besitzt also keinen natürlichen Ursprung
- Um die Item- und Personenparameter schätzen zu können, muss eine Normierung durchgeführt werden, d.h. ein Nullpunkt festgelegt werden

2.4 Normierung & Messniveau

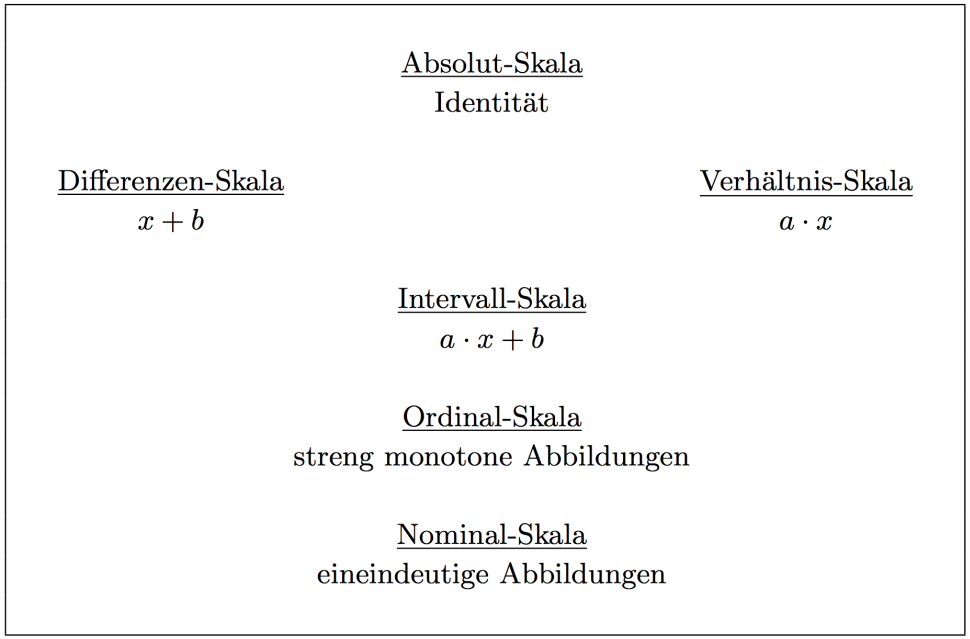
- Häufig wird die Summe (und damit der Mittelwert) der Itemschwierigkeiten als Nullpunkt der Skala festgelegt (*Summennormierung*)

$$\sum_{j=1}^{m} \beta_j = 0$$

  - Items, die leichter (schwerer) als der Mittelwert der untersuchten Items sind, erhalten einen negativen (positiven) Itemparameter
- Mit der Normierung der Itemparameter sind auch die Personenparameter eindeutig festgelegt

2.4 Normierung & Messniveau

- Welches Messniveau hat das Rasch Modell?



- Das Rasch Modell erlaubt mindestens intervallskalierte Messungen; manche argumentieren, dass es Messungen auf Differenz-Skalenniveau erlaubt

Abb. aus Strobl (2012)

Item Response Theorie

1. Hinführung zum Rasch Modell
    1. Summenscore
    2. Lösungswahrscheinlichkeit
2. Rasch Modell
    1. Modellgleichung
    2. Item Characteristic Curve
    3. Annahmen und Eigenschaften
    4. Normierung & Messniveau
    5. **Maximum Likelihood Schätzung**
3. Beispiel Algebratest

2.5 Maximum Likelihood Schätzung

- Ziel: Schätzung der Item- und Personenparameter aus den Daten
- Mit der Maximum Likelihood (ML) Schätzung werden die Modellparameter so geschätzt, dass die Plausibilität (likelihood) der beobachteten Daten maximiert wird.
- Dazu wird das Maximum der Likelihood Funktion gesucht
- Die Likelihood Funktion gibt an, wie wahrscheinlich es ist, die beobachteten Daten zu erhalten, wenn in der Population die geschätzten Parameter gelten würden

$$L_{u_{ij}}(\theta_i, \beta_j) = P(u_{ij} \mid \theta_i, \beta_j) = \frac{e^{u_{ij} \cdot (\theta_i - \beta_j)}}{1 + e^{\theta_i - \beta_j}}$$

2.5 Maximum Likelihood Schätzung

- Für die ML-Schätzung der Modellparameter wird ein iteratives Verfahren eingesetzt
  - Einsetzen von Startwerten für die Modellparameter
  - Bestimmung der Likelihood Funktion
  - Einsetzen von neuen, besseren Werten für die Modellparameter
  - Bestimmung der Likelihood Funktion
  - usw.
  - Nach jedem Schritt wird die Veränderung zum vorherigen Schritt registriert
- Die Schätzung wird beendet, wenn ein Abbruchkriterium erfüllt ist
  - Maximale Anzahl von Iterationen
  - Differenz der Schätzungen aus zwei aufeinanderfolgenden Iterationen ist geringer als ein Genauigkeitskriterium

Item Response Theorie

1. Hinführung zum Rasch Modell
    1. Summenscore
    2. Lösungswahrscheinlichkeit
2. Rasch Modell
    1. Modellgleichung
    2. Item Characteristic Curve
    3. Annahmen und Eigenschaften
    4. Normierung & Messniveau
    5. Maximum Likelihood Schätzung
3. **Beispiel Algebratest**

3. Beispiel: Algebratest

- *N* = 715 Studierende (69% weiblich)
- Algebratest aus 9 Items

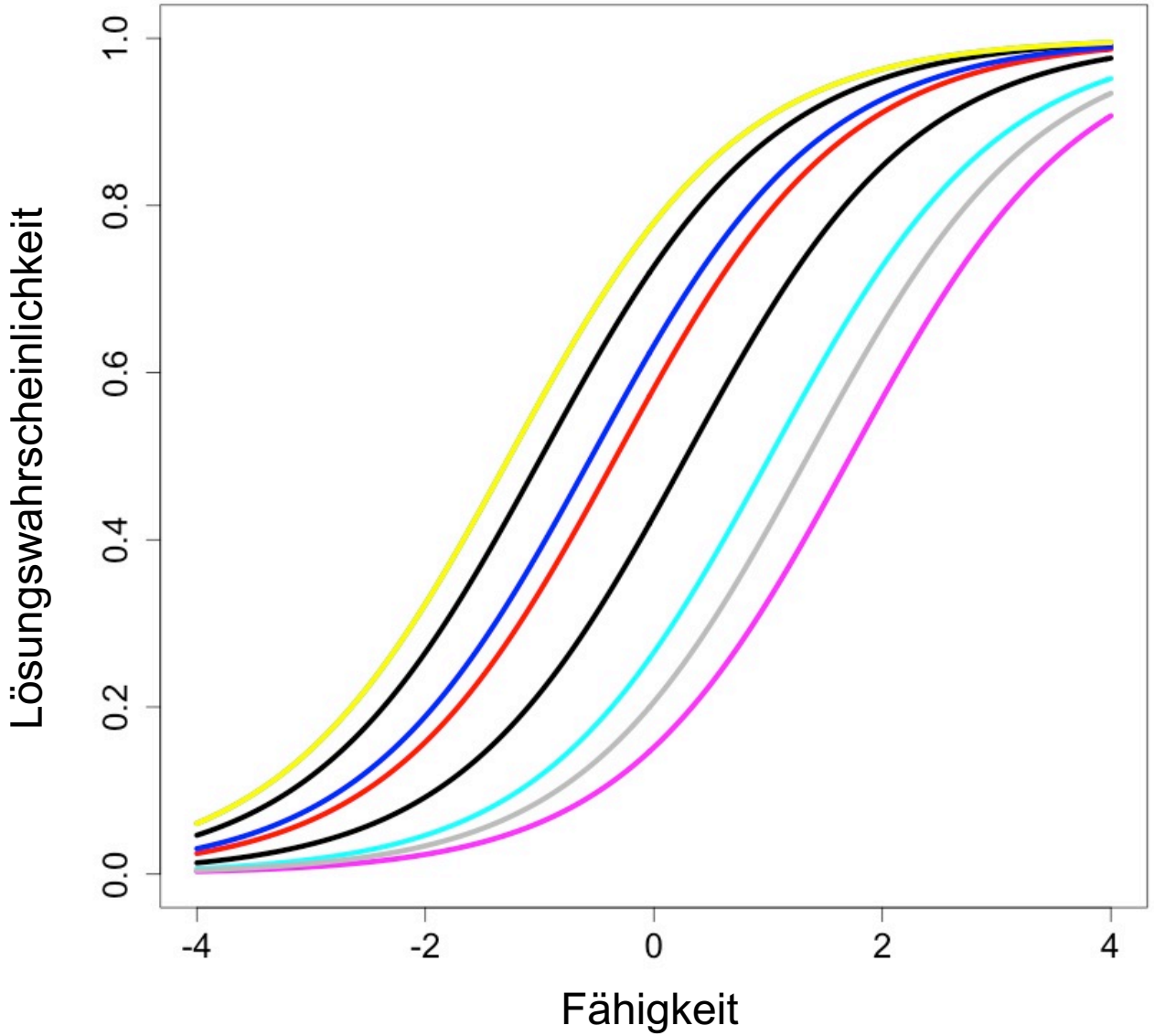
$-16 + 7 - 9$ =	-18	18	0	?	20	ITEM1
$3\frac{3}{7} - \frac{8}{7}$ =	$-3\frac{5}{7}$	$\frac{1}{7}$	$2\frac{2}{7}$	?	21	ITEM2
$9 - (-4 + 7)$ =	6	12	20	?	22	ITEM3
9% von 270 sind	30	24,3	3	?	23	ITEM4
$\lvert -(17 - 4) \rvert$ =	-13	13	21	?	24	ITEM5
3^{-2} =	$\sqrt{3}$	9	$\frac{1}{9}$	?	25	ITEM6
$5 - 6*7$ =	-37	-7	37	?	26	ITEM7
$(x^a)^b$ =	x^{a+b}	$x^{(ab)}$	x^{a*b}	?	27	ITEM8
$x \leq -5$ ist falsch für	$x = -5$	$x = -3$	$x = -12$	?	28	ITEM9

3. Beispiel: Algebratest

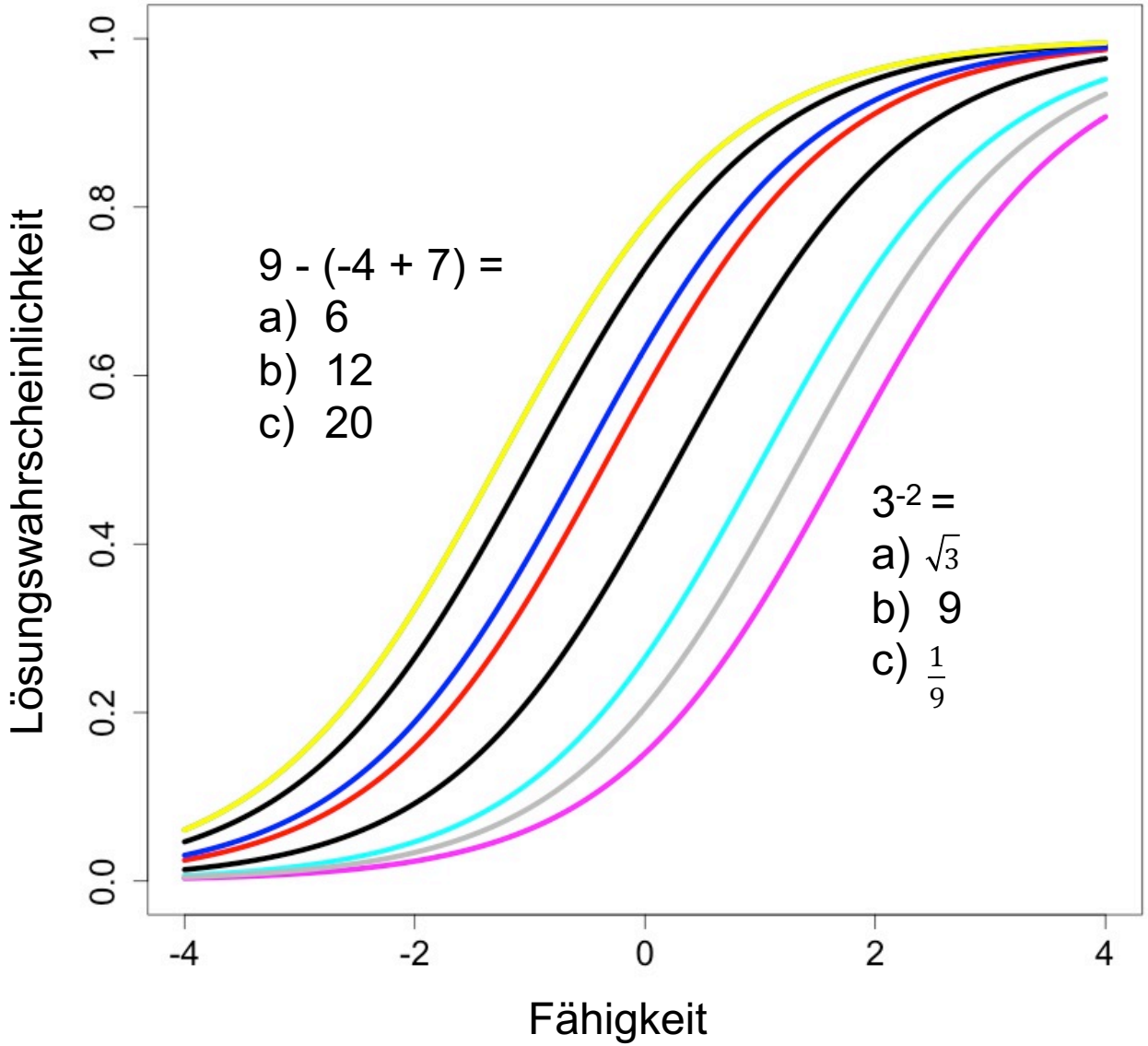
- Das Rasch Modell wird auf die Daten angepasst

Item	Itemschwierigkeit
1	-0.98
2	-0.33
3	-1.26
4	-0.54
5	1.02
6	1.72
7	-1.26
8	1.34
9	0.29

3. Beispiel: Algebratest



3. Beispiel: Algebratest



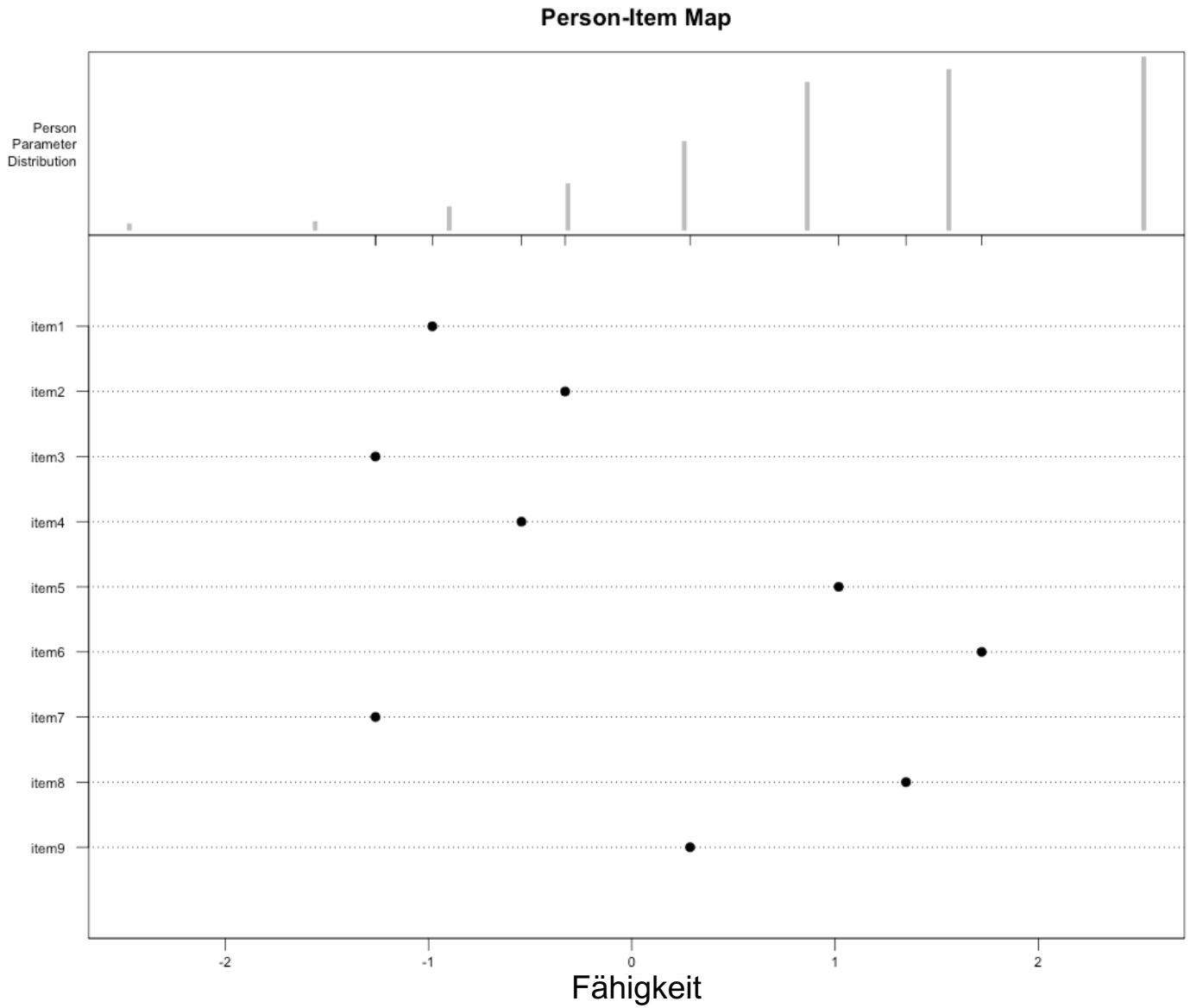
3. Beispiel: Algebratest

Fall	Summen-score	Max. Summen-score	Personen-parameter	SE Personen-parameter
1	8.00	9.00	2.20986	1.02769
2	8.00	9.00	2.20986	1.02769
3	9.00	9.00	3.54575	1.64382
4	6.00	9.00	0.80090	0.79482
5	8.00	9.00	2.20986	1.02769
6	6.00	9.00	0.80090	0.79482
7	5.00	9.00	0.23281	0.76253
8	7.00	9.00	1.43297	0.86334
9	6.00	9.00	0.80090	0.79482
10	5.00	9.00	0.23281	0.76253
11	7.00	9.00	1.43297	0.86334
12	6.00	9.00	0.80090	0.79482
13	7.00	9.00	1.43297	0.86334
14	6.00	9.00	0.80090	0.79482
15	8.00	9.00	2.20986	1.02769
16	3.00	9.00	-0.84698	0.77567
17	9.00	9.00	3.54575	1.64382
18	8.00	9.00	2.20986	1.02769

Ausgabe der Personenparameter

Fälle mit gleichen Summenscores erhalten auch gleiche Fähigkeitsschätzer → suffiziente Statistiken

3. Beispiel: Algebratest



Verteilung der Personen-parameter

Itemschwierig-keiten

Literatur zu dieser Sitzung

Strobl (2012). *Das Rasch-Modell: Eine verständliche Einführung für Studium und Praxis*. München: Rainer Hampp Verlag.
Kapitel 2 (ohne Herleitungen)