Institut zur Qualitätsentwicklung im Bildungswesen

Übung: einfache IRT-Modelle in R

Sebastian Weirich und Nicklas Hafiz

Institut zur Qualitätsentwicklung im Bildungswesen (IQB)
Humboldt-Universität zu Berlin

Gesis Workshop, Oktober 2024

Überblick: R-Pakete für IRT-Modellierung

- Ca. 45 R-Pakete für IRT-Modellierung (Choi & Asilkalkan, 2019)
- In diesem Workshop werden jedoch nur vier Pakete betrachtet bzw. verwendet

Name	Autoren	Ort	Installation	Features und Vorteile	Nachteile
TAM	Alexander Robitzsch, Thomas Kiefer, Margaret Wu	CRAN	install.packages ("TAM")	Vielfältige Modelle, sehr schnell, sehr flexibel, plausible values	eingeschränkt einsteigerfreundlich
lme4	Douglas Bates et al.	CRAN	install.packages ("lme4")	große Flexibilität bei Modellspezifikation; instruktiv für das Verständnis der IRT	kein originäres IRT-Paket, teils langsam, nur Modelle aus der 1PL-“Familie“
mirt	Phil Chalmers et al.	CRAN	install.packages ("mirt")	sehr flexibel, auch 2pl, 3pl, mixed IRT, plausible values	Modelle sind teils anspruchsvoll zu spezifizieren
eatModel	Sebastian Weirich, Karoline Sachse, Benjamin Becker	Github	remotes::install_ github("weirichs/ eatModel", upgrade= "never")	Einsteigerfreundlich, Konsistenzprüfungen	weniger flexibel, weniger schnell, nicht sonderlich effizient programmiert, nicht auf CRAN verfügbar

Erste Übung

- Bitte dazu das Skript

  Tag1_2Nachmittag_Nr1_einfache_IRT_Modelle.r

  öffnen

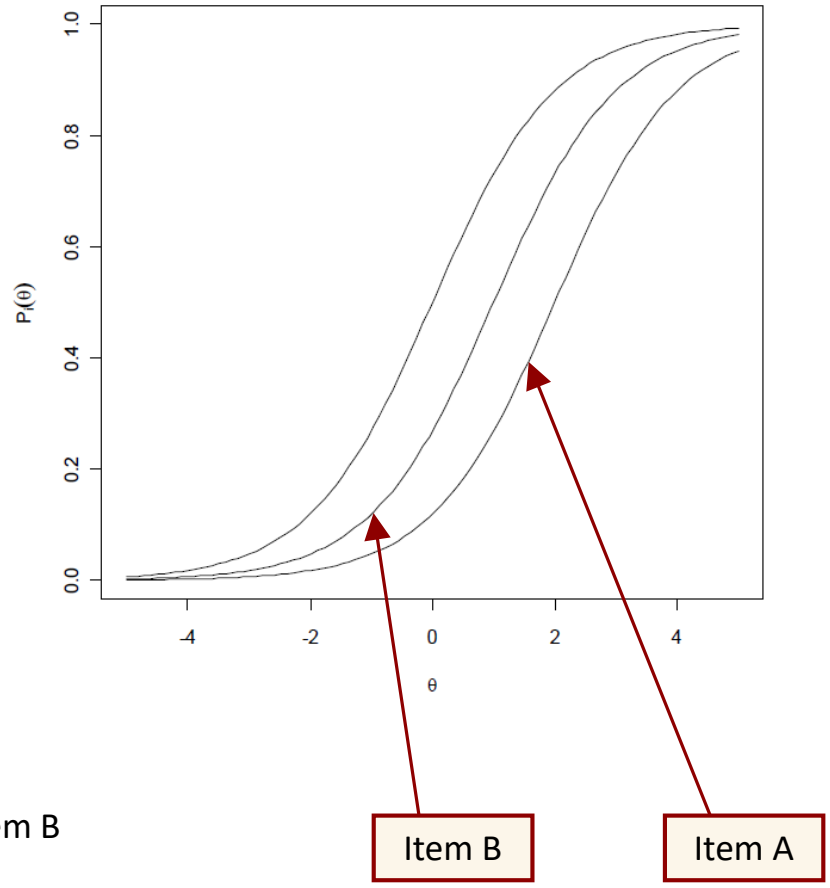
```r
# 0. Vorbereitung: benoetigte Pakete installieren
#######################################################

# schauen, ob das Paket installiert ist
# dazu erstmal alle installierten pakete auflisten
allPackages <- installed.packages()

# welche der benoetigten Pakete sind in welcher Version vorhanden?
allPackages[grep(c("TAM|lme4|mirt|eatModel"),allPackages[,"Package"]),c("Package", "Version")]

# das fehlende bei Bedarf installieren
install.packages("TAM")
install.packages("lme4")
install.packages("mirt")
remotes::install_github("weirichs/eatModel", upgrade= "never")

```

- Nächster Schritt: Prüfen der drei Voraussetzungen des Raschmodells
  1. Voraussetzung: Raschhomogenität (parallele Item-Response-Kurven)

Raschmodell, Annahme 1: parallele Itemcharakteristikkurven

$$\text{logit}\left(P(X_{ni}=1)\right)=\theta_n-\beta_i$$

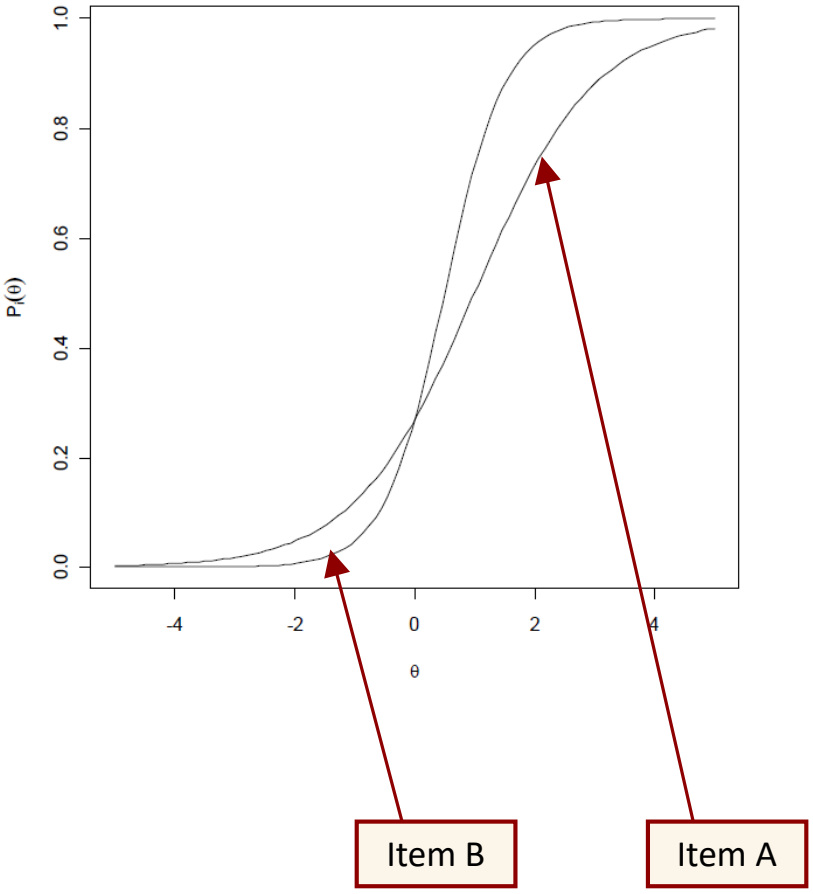
- θ_n: unidimensional latent trait
- Doppelte Monotonizität
  - Rangfolge der Items ist für alle Personenpopulationen gleich
  - Rangfolge der Personen ist für alle Itempopulationen gleich
  - Beide Annahmen folgen aus der Annahme paralleler Itemcharakteristikkurven (ICC) im Raschmodell (gleiche Trennschärfe für alle Items)
- Parallele Itemcharakteristikkurven
  - Kurven überschneiden sich nicht
  - Item A ist für jede beliebige Person und in jeder beliebigen Population schwerer als Item B



Alternativ: 2PL-Modell, keine parallelen ICCs

$$\text{logit}\left(P(X_{ni}=1)\right)=\theta_n-\alpha_i\beta_i$$

- Item A ist schwerer als Item B
- Item B hat eine höhere Trennschärfe als Item A
- Für weniger fähige Personen hat Item B eine geringere Lösungswahrscheinlichkeit als Item A **(obwohl Item B das leichtere Item ist)**
- Für fähige Pesonen hat Item B eine höhere Lösungswahrscheinlichkeit als Item A
  - ggf. schwere Interpretierbarkeit
  - In einem 2PL-Modell wäre es bspw. schwierig, Kompetenzstufen für Personen und Items zu definieren, die beidemale dieselbe Intervallbreite (z.B. 75 Punkte) haben

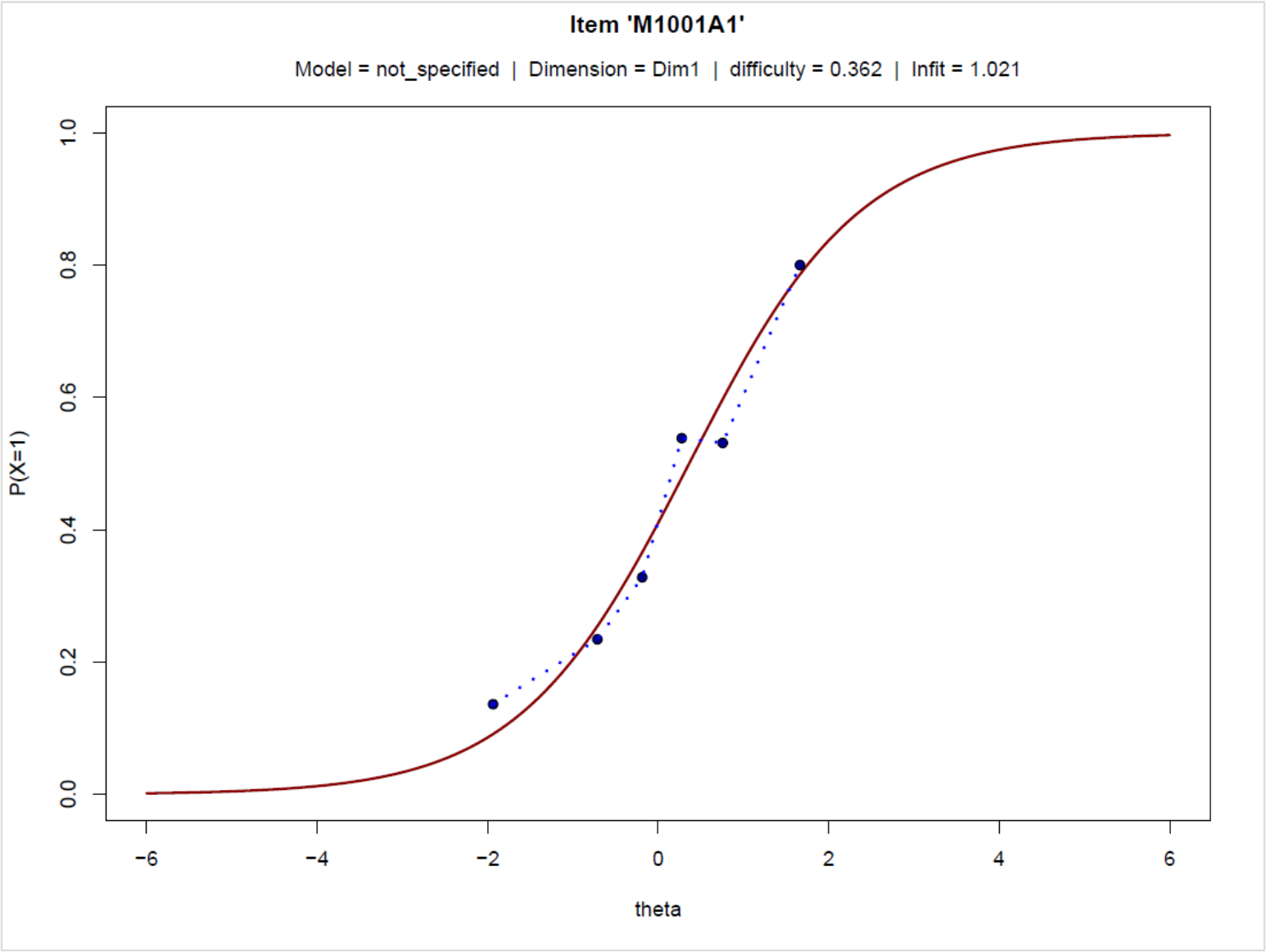


Prüfung der Raschhomogenität

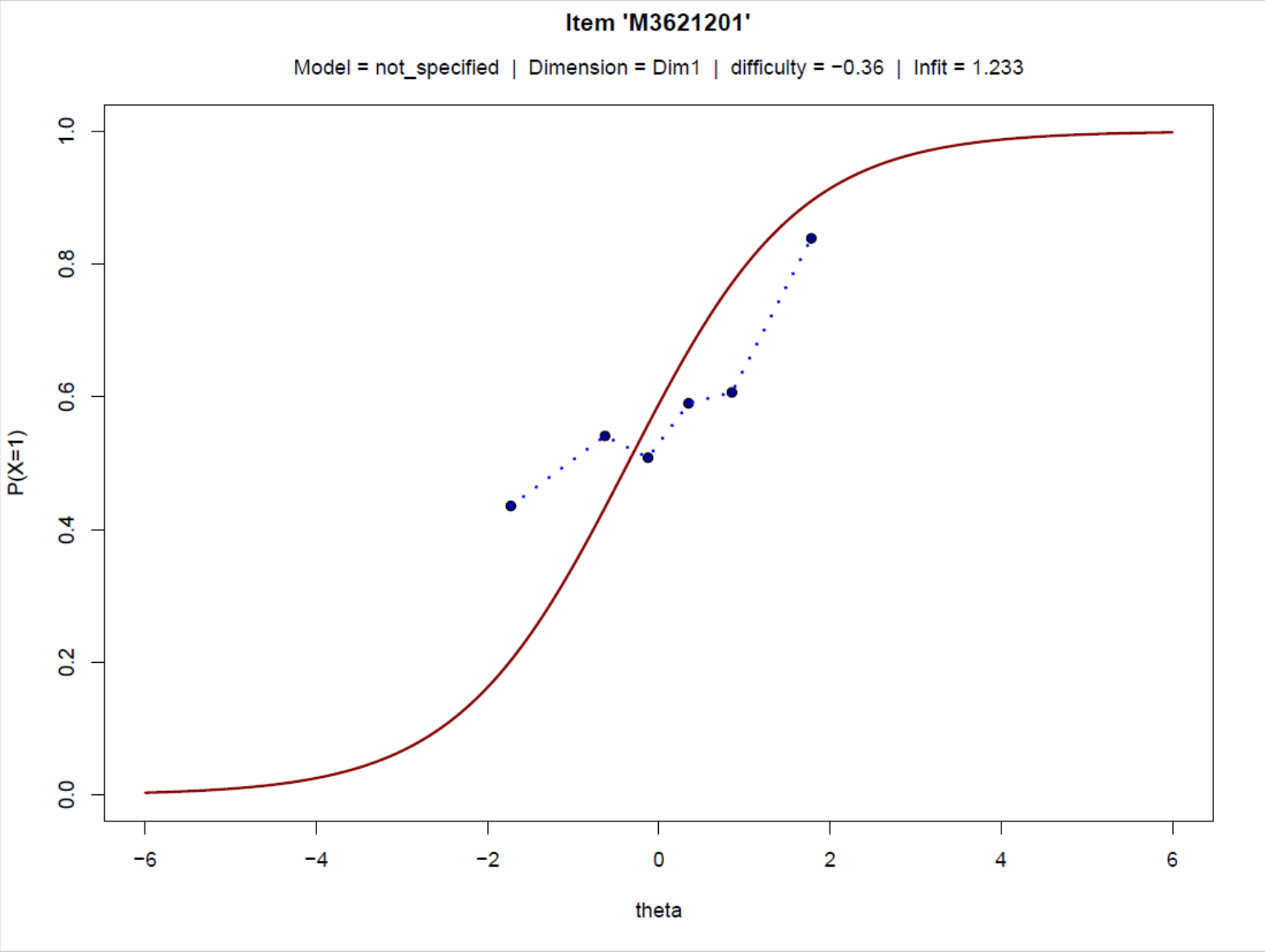
1. Variante: Itemfit (Infit)

- Beruht auf einem Vergleich der empirischen Item-response-Kurve mit der durch das Raschmodell implizierten Item-Response-Kurve
- Idealerweise sollte der Infit = 1 sein
- Werte < 1, Overfit: die empirische Kurve verläuft steiler als die modellimplizierte
- Werte > 1: die empirische Kurve verläuft flacher als die modellimplizierte Kurve. Werte > 1.15 gelten in der Regel als kritisch (zuweilen findet man aber auch 1.25 oder sogar 1.5 als Grenze)
- Möglichkeit: Plotten der itemspezifischen Response-Kurven

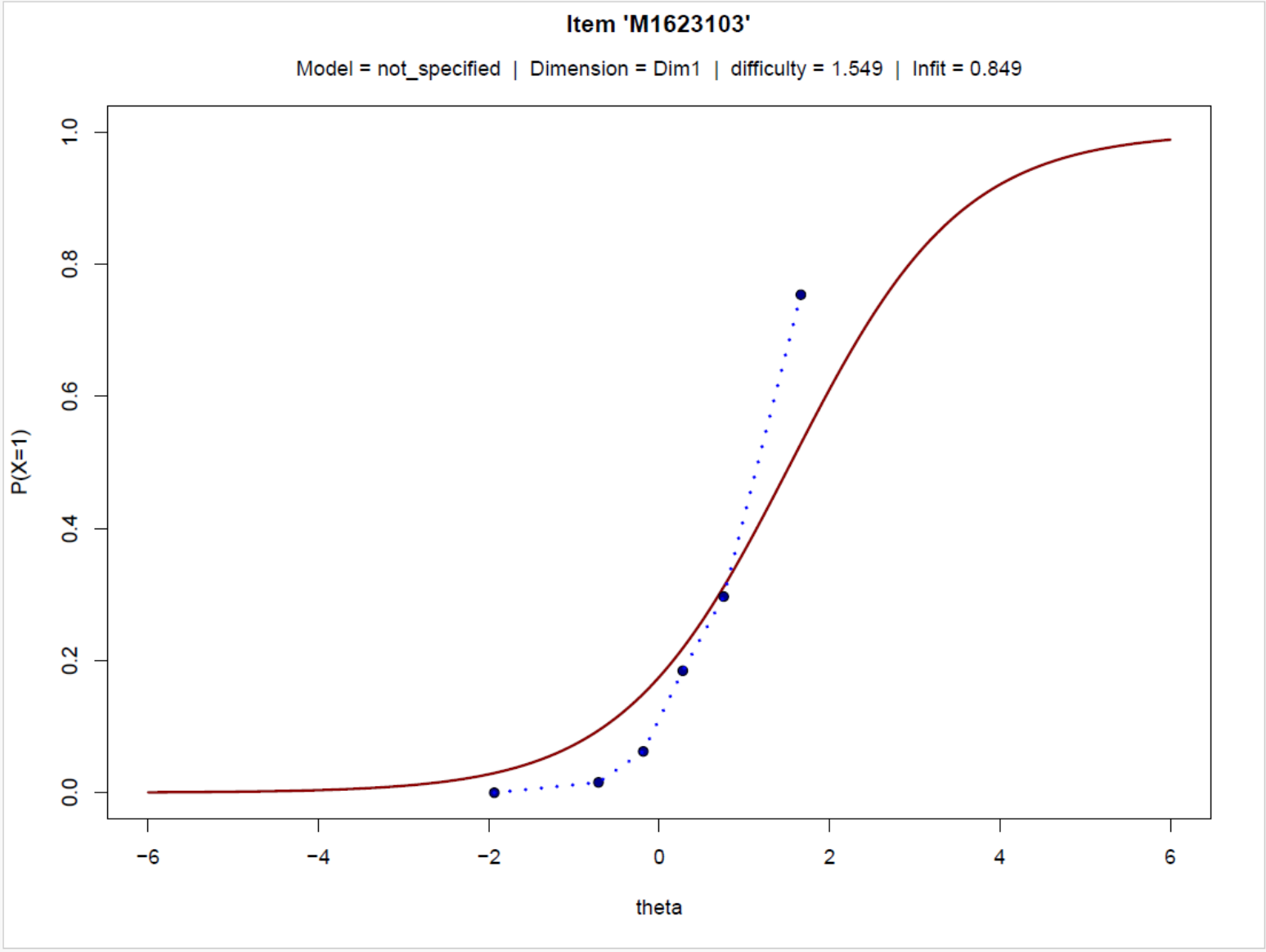
Infit: Beispiel für guten Fit



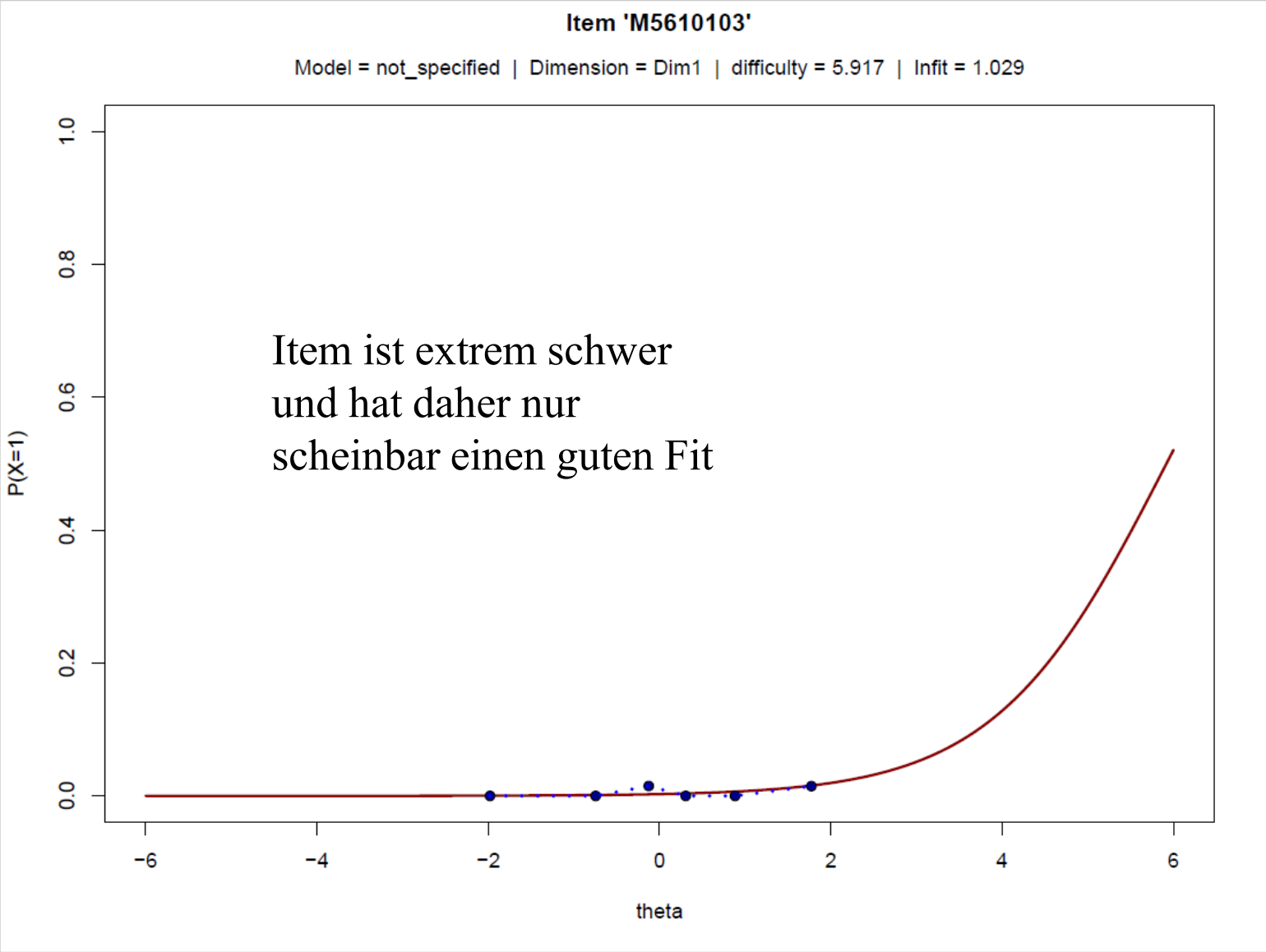
Infit: Beispiel für schlechten Fit



Infit: Beispiel für “Overfit”



Aber: Infit kann missverständlich sein



Prüfung der Raschhomogenität

2. Variante: Vergleich gegen ein zweiparametrisches IRT-Modell (2PL- oder Birnbaum-Modell)

- Spezifizieren zweier “konkurrierender” Modelle und anschließender Test, welches Modell die empirischen Daten besser beschreibt
  - Genauer: ist die Wahrscheinlichkeit der empirischen Daten unter der 1PL-Modellannahme oder unter der 2PL-Modellannahme größer?
  - Noch genauer: die Wahrscheinlichkeit der empirischen Daten ist unter Annahme eines liberaleren Modells (sofern das strengere Modell in dem liberaleren genestet ist) immer größer
  - Ist die Modellpassung so viel besser, dass sie die Spezifizierung der zusätzlichen Modellparameter rechtfertigt?

Prüfung der lokalen stochastischen Unabhängigkeit

- Q3-Statistik von Yen (1984, 1993)
  - Residualkorrelationen von Itempaaren sollten 0 sein
  - Abweichungen nach oben und unten von ± 0,25 in der Regel akzeptabel

Eigene Übung

- Empirischer IRT-Datensatzes aus der Evaluation der Bildungsstandards (trends)
  - zwei Kompetenzbereiche „reading“ und „listening“
  - Kohortenvergleich mit drei Messzeitpunkten (2010, 2015, 2020)
  - Personen stammen aus drei Ländern (anonymisiert)
  - Dichotome Items (0/1)
  - Itemformate: offen, geschlossen (multiple choice), halb offen

Datensatz im Langformat vs. Wideformat

- Wideformat
  - Eine Zeile pro Person
  - Datensatz kann nicht gleichzeitig Eigenschaften der Personen und Eigenschaften der Items abbilden

ID	Geschlecht	Item 1	Item 2	Item 3
1	weiblich	0	1	1
2	weiblich	0	0	1
3	männlich	0	0	1

- Langformat
  - Eine Zeile pro Beobachtung (Person $\times$ Item Kombination)
  - Datensatz kann gleichzeitig Eigenschaften der Personen und Eigenschaften der Items abbilden
  - Für eine IRT-Modellierung muss der Datensatz gegebenenfalls ins Wideformat umgeformt werden

ID	Geschlecht	Item	Format	Domain	Response
1	weiblich	1	MC	reading	0
2	weiblich	1	MC	reading	0
3	männlich	1	MC	reading	0
1	weiblich	2	halboffen	listening	1
2	weiblich	2	halboffen	listening	0
3	männlich	2	halboffen	listening	0
1	weiblich	3	offen	listening	1
2	weiblich	3	offen	listening	1
3	männlich	3	offen	listening	1

Eigene Übung

- Prüfen Sie für den Teildatensatz des Jahres 2010
  - Ob die Items rasch-homogen sind bzw. einen akzeptablen Fit haben
  - Ob eher 1pl oder 2pl Modellierung angeraten ist
  - Ob die Items lokal stochastisch unabhängig sind

```
# empirischen Uebungsdatensatz aufbereiten
# Datensatz enthaelt Item- UND Personeninformationen
data(trends)

# Personendatensatz ins wideformat transformieren
dat_wide <- tidyr::pivot_wider(subset(trends, year == 2010),
            id.cols = c("idstud", "sex", "ses", "ses"), names_from = "item",
            values_from = "value") %>% as.data.frame ()

# Iteminformationen als separates Objekt
item_info <- unique(subset(trends, year == 2010)[,c("item", "domain", "format")])

# Q Matrix enthaelt Zuordnung der Items zu Dimensionen
qmat <- data.frame(item_info, model.matrix(~domain - 1, data = item_info))
```

Annahme 2, Unidimensionalitätsannahme

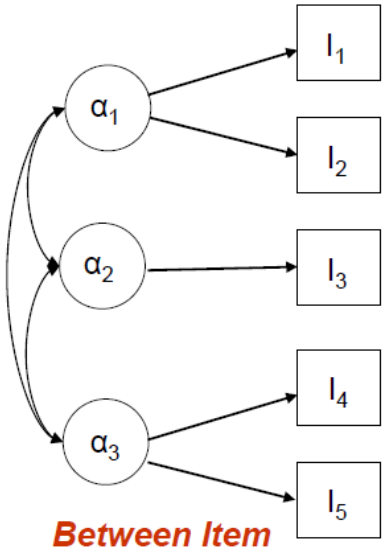
- Die Wahrscheinlichkeit $P(X_{ni} = 1)$ wird lediglich durch θ_n und β_i bestimmt
- Prüfung erfolgt indirekt
  - Man testet Annahmen, die aus dieser Unidimensionalität resultieren
  - Anders gesagt: was wären mögliche Konsequenzen, wenn die Items eines Tests *nicht* eindimensional wären?
    - Invarianz verletzt: differentielles Itemfunktionieren (differential item functioning; DIF)
    - Mehrdimensionale IRT-Modelle
    - Kontexteffekte
    - …
- Differential Item Functioning (DIF): ist der Test fair?
  - DIF-Modell: $\text{logit}\left(P(X_{ni}=1)\right)=\theta_n-\beta_i+\tau_1 g_j+\tau_2\left(\beta_i g_j\right)$
  - Test des Interaktionsterms $\tau_2\left(\beta_i g_j\right)$
  - formal: g_j ist ein Indikator für die Gruppe, τ ist der (Haupt-)Effekt der Gruppe
  - Die Interaktion beschreibt, ob ich zu einem anderen Gruppeneffekt kommen würde, wenn ich andere Testitems verwenden würde … das wäre DIF und der Test damit potenziell unfair

DIF: Differential Item Functioning

- DIF ist das “Gegenteil” von Messinvarianz
  - DIF bedeutet, dass die Messeigenschaften eines Items sich bspw. zwischen Gruppen (männlich, weiblich; deutsche Muttersprache, nicht-deutsche Muttersprache) unterscheiden
  - Test wäre dann im Extremfall nicht mehr fair
  - Bsp.: Mathematiktest, der sprachlich anspruchsvolle Aufgabenformulierungen enthält und daher Personen nicht-deutscher Muttersprache benachteiligt: obwohl deren „wahre“ mathematische Kompetenz genau so groß wäre, würden sie schlechter abschneiden, als Personen deutscher Muttersprache

Mehrdimensionale IRT-Modelle

- Konfirmatorische Spezifizierung der Mehrdimensionalität
- Vergleich zweier konkurrierender Modelle (ein- vs. mehrdimensional; vergleichbar des Vergleichs 1pl vs. 2pl)



Between Item Dimensionality

- Q-Matrix beschreibt Zuordnung von Items zu Attributen (Skills, Traits)

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$