

Fragebogen 6

Testtheorie & Testkonstruktion

Ausarbeitung: Johannes Geffers [auf Grundlage einiger Vorarbeiten...]

1. Worin unterscheiden sich die Item Response Modell von Rasch und Birnbaum?

- Das RASCH-Modell ist ein **Spezialfall** des BIRNBAUM-Modells.
 - Diskriminationsparameter konstant
 - Rateparameter $c_i=0$.

2. Zwischen welchen Größen stellt im Rahmen der probabilistischen Testmodelle eine Itemcharakteristik einen Zusammenhang her, und welche Größen werden in einer Testcharakteristik miteinander verbunden?

- Die Itemcharakteristik stellt eine Beziehung her zwischen der **Lösungswahrscheinlichkeit** einer Aufgabe und der **Personenfähigkeit**.
- Diese Beziehung kann mathematisch beschrieben werden durch eine logistische Funktion, die als Parameter die Personenfähigkeit einerseits und die Itemparameter andererseits enthält:
- ◆ **Itemparameter** im Fall des drei (bzw. zwei-)parametrischen BIRNBAUM-Modells:
 - **Schwierigkeitsparameter b**: Entspricht der Lokation des Wendepunkts der Kurve auf der Fähigkeitsskala (Abszisse). Dort beträgt die Lösungswahrscheinlichkeit 0.5, bzw. Fähigkeit = Schwierigkeit.
 - **Diskriminationsparameter a**: Entspricht der Steigung der Kurve am Wendepunkt; beschreibt, wie gut das Item zwischen Personen mit unterschiedlichen Fähigkeiten diskriminiert.
 - **Rateparameter c**: Ist die untere Asymptote der ICC (item-characteristic-curve); beschreibt die Wahrscheinlichkeit eines Probanden ohne jegliches Wissen, das Item dennoch richtig zu beantworten.
- Itemparameter im Fall des zwei-parametrischen BIRNBAUM-Modells:
 - Rateparameter entfällt
- RASCH-Modell.
 - Rateparameter = 0
 - Diskriminationsparameter konstant

3. Zeigen Sie, dass die Lösungswahrscheinlichkeiten für eine Person, die so fähig ist wie das betrachtete Item schwierig, im Rasch und im Birnbaum-Modell 0.50 beträgt! Wie groß ist diese Wahrscheinlichkeit, wenn zusätzlich noch ein Rateparameter in das Modell aufgenommen wird?

➤ Vorausgesetzt: $\Theta = \text{Schwierigkeitsparameter } b_i$

- RASCH (1-Parameter-Modell).

$$P_i(\Theta) = \frac{e^{(\Theta - b_i)}}{[1 + e^{(\Theta - b_i)}]} = \frac{e^0}{[1 + e^0]} = \frac{1}{[1 + 1]} = 0.5$$

- BIRNBAUM, mit Rateparameter c_i .

$$P_i(\Theta) = c_i + (1 - c_i) \cdot \left[\frac{1}{[1 + e^{-Da_i(\Theta - b_i)}]} \right] = c_i + (1 - c_i) \cdot \left[\frac{1}{[1 + e^0]} \right]$$

$$P_i(\Theta) = c_i + (1 - c_i) \cdot 0.5 = c_i + 0.5 - 0.5 c_i = 0.5 + 0.5 \cdot c_i \quad \bullet \quad \text{Mit } c_i = 0 :$$

$$P_i(\Theta) = 0.5 + 0.5 \cdot c_i = 0.5 + 0.5 \cdot 0 = 0.5$$

4. Was ist unter einer Iteminformationsfunktion zu verstehen, in welcher Beziehung steht sie zur Itemcharakteristik, und an welcher Stelle hat sie ihr Maximum? Welche Anhaltspunkte kann man der Iteminformationsfunktion für die Beurteilung der Brauchbarkeit eines Items entnehmen?

- Die **Iteminformationsfunktion** beschreibt das Verhältnis zwischen dem Informationsgehalt (informativeness) eines Items und dem Personenparameter θ . Die Iteminformationsfunktion eines bestimmten θ -Wertes ist mathematisch das Verhältnis zwischen der Steigung der ICC und dem zu erwartenden Messfehler dieses θ -Wertes.

$$I(\theta, u_i) = \frac{[P_i'(\theta)]^2}{[P_i(\theta)Q(\theta)]}$$

$P_i'(\theta)$ entspricht der ersten Ableitung (der Steigung) der ICC am Punkt θ

$Q(\theta)$ ist die Wahrscheinlichkeit einer falschen Antwort, oder $1 - P_i(\theta)$

- Grundregel: Je **steiler die Steigung der ICC, desto besser ist die diskriminierende Eigenschaft des Items** und desto höher ist der Wert des Parameters.
- Mit der Zunahme der Steigung wird die Bandbreite der Fähigkeitswerte zwischen denen diskriminiert wird kleiner: Das **Bandbreiten-Paradoxon** beschreibt das ›Tauschgeschäft‹ zwischen der Bandbreite der zu erfassenden Fähigkeitswerte zwischen denen diskriminiert wird und der Steile des Anstiegs der Steigung am Wendepunkt (a-Parameter) der Itemcharakteristik-kurve (und auch in den ›Randbereichen‹ der Kurve – eine große Steigung im Mittelbereich bedeutet eine geringere Steigung [und damit Diskriminationsfähigkeit] in den Randbereichen).
- Der **Informationsgehalt eines Items** wird zudem durch den **Messfehler** beeinflusst, der für dieses Item zu erwarten ist. In der klassischen Testtheorie wird ein Standardmessfehler für alle Items angenommen, was aber nicht der Realität entspricht. Je kleiner der zu erwartende Messfehler an einem bestimmten Punkt θ , desto höher ist der Informationsgehalt des Items an diesem Punkt.
- Ihr **Maximum** hat die Iteminformationsfunktion an dem Punkt, wo das Verhältnis von Informationsgehalt des Items zu dessen Genauigkeit am größten ist (siehe obige Formel?).
- Testinformationsfunktion (test information function)**. Der Informationsgehalt eines Tests an einem bestimmten Punkt θ ist die Summe aller Iteminformationsfunktionen für diesen θ -Wert:

$$I(\theta) = \sum_{i=1}^N P_i'(\theta)^2 / P_i(\theta)Q(\theta)$$

- **Je größer die Testinformationsfunktion an einem bestimmten Punkt θ ist, desto besser diskriminiert der Test an diesem Punkt und desto niedriger ist dort sein Messfehler.**
- Unterschied zur KTT: Der **Standardmessfehler** ist hier keine Statistik, sondern eine **Funktion des Fähigkeitswertes θ !**
- Für die **Reliabilität** bedeutet dies: **Kein Test ist für alle Personen** (bzw. Fähigkeitsstufen oder -grade) **gleich reliabel** – vielmehr hat jeder Test eine **Bandbreite von Fähigkeitswerten, für die er am geeignetsten ist!**

5. Berechnen sie im Birnbaum-Modell die Lösungswahrscheinlichkeiten für eine Person mit einem Fähigkeitswert von 1.00 für folgende Items:

- Item 1** mit einer Schwierigkeit von -1 und einer Diskriminationsfähigkeit von **0.85**
- Item 2** mit einer Schwierigkeit von 0.5 und einer Diskriminationsfähigkeit von **1.3** und
- Item 3** mit einer Schwierigkeit von 2 und einer Diskriminationsfähigkeit von **1.7?**

[Rechenaufgabe, deshalb nicht ausgeführt.]

6. Worin besteht der Grundgedanke des Linearen Logistischen Testmodells (LLTM), und in welchen Bereichen und unter welchen Voraussetzungen ist dieses Modell vor allem einsetzbar?

- Das LLTM ist eine Erweiterung des einfachen RASCH-Modells.

$$p(+; v, i) = \frac{e^{(\theta_v - \sigma_i)}}{[1 + e^{(\theta_v - \sigma_i)}]} \quad (\text{einfaches RASCH-Modell}) \quad (1)$$

- Der **Grundgedanke** des LLTM besteht darin, die **Item-Schwierigkeiten in elementare Komponenten zu zerlegen**, denen jeweils wieder eine bestimmte Schwierigkeit zukommt; die einzelnen Komponenten werden dann zu einer linearen Kombination (wieder) zusammengefügt.

- **Dekomposition der Itemschwierigkeit** in eine lineare Kombination elementarer Komponenten:

$$\sigma_i = \sum_{j=1}^m q_{ij} \eta_j + c \quad (2)$$

η_j ($j = 1, 2, \dots, m$) Parameter für die elementaren Komponenten

q_{ij} hypothetische Frequenzen, mit der jede Komponente j die Lösung des Items i beeinflusst

c Skalierungskomponente

- **Elementare Komponenten**: Beispielsweise *kognitive Operationen* (charakterisiert durch ihre jeweilige Schwierigkeit) die notwendig für die Lösung eines Items sind oder *Instruktionsbedingungen* (charakterisiert durch ihre Effizienz), die Probanden erhalten, bevor sie versuchen ein Item zu lösen.

- **Annahmen und Voraussetzungen.**

- LLTM ist beschränkt auf Tests, für die **elementare und stabile Itemparameter identifiziert und erfolgreich bei der Konstruktion neuer Items angewendet** werden können.
 - Dies trifft wahrscheinlich eher auf Mathematik- und Physik-Tests zu, denn für Tests in Geschichte, Soziologie oder auch Persönlichkeitstests.
- Alle Personen müssen den **gleichen Lösungsalgorithmus** benutzen.
 - Dies schließt die Möglichkeit aus, dass unterschiedliche Personen aufgrund einer unterschiedlichen Lerngeschichte Items auf unterschiedlichen Wege richtig lösen.
- **Veränderungen werden nur als quantitative begriffen.**
 - Dies schließt die Möglichkeit einer qualitativen Modellierung von entwicklungsbedingten Veränderungen aus.
- Diese Forderungen stehen im Widerspruch zu fast allen psychologischen Entwicklungsmodellen!

- **Verwendung.**

- ◆ Die LLTM wird genutzt, um die **Testleistung auf die Fähigkeit** zu beziehen, bestimmte kognitive Operationen auszuführen (u.a. *PIAGET-Test*). Auf Grundlage eines *global logistic test modell* wurden beispielsweise vier Testmodelle entwickelt, die unterschiedliche Typen von Lerneffekten berücksichtigen sollen (*globale Lerneffekte*, *personenspezifische Lerneffekte*, *itemspezifische Lerneffekte* und *globale Lerneffekte bei heterogenem Itemsample*).
- ◆ Die LLTM bringt auch Vorteile in der **Testentwicklung**:
 - Testkonstrukteure werden dazu verpflichtet, eine klare Aufgabenstruktur zu erstellen, bzw. zu diese zu analysieren.
 - Kontentvalidität kann erreicht werden, wenn der Test aufgebaut ist auf einer repräsentativen Stichprobe aus einer Domäne der Items, innerhalb derer alle Items objektiv mittels spezifischer Konstruktionsregeln konstruiert werden.
 - Adaptives Testen: Die LLTM ermöglicht die systematische Konstruktion von Items mit vorhersagbaren Schwierigkeiten für Item(daten-)banken. (Wegen Lerneffekten bei der Durchführung des Tests ist dies jedoch nicht ganz unproblematisch.)

7. Für die Beantwortung welcher Fragen, für die Lösung welcher Probleme sind die probabilistischen Modelle vor allem interessant? Wo liegen die Grenzen ihrer Anwendbarkeit?