Psychometrie

Lena Straub

02.12.2011

1 Allegmeine Beschreibung

Die Psychometrie beschäftigt sich damit, mathematische bzw. statistische Methoden zu entwickeln, um psychologische Messungen zu analysieren. Psychologische Eigenschaften können nicht direkt gemessen werden, man bezeichnet sie als latente Variablen. Stattdessen werden manifeste Variablen gemessen, von denen man dann auf die latenten schließen will. Man kann mit der Psychometrie die Unterschiede zwischen Personen analysieren.
Die Psychometrie ist seit der Gründung der „Psychometric Society"1935 eine „formal discipline" *(Michael W. Browne, Journal of the American Statistical Association, S.661)*. Diese Society brachte die Zeitschrift „Psychometrika" heraus mit dem Ziel, die Psychologie als quantitative Wissenschaft zu etablieren.
Man kann Psychometrie in drei Teilgebiete aufteilen:

- Theorie psychologischer Tests
- Faktorenanalyse und dazugehörige Modelle
- Multidimensionale Skalierung

1.1 Theorie psychologischer Tests

Bei der Theorie psychologischer Tests geht es darum, Aufgaben oder Items zu erstellen, deren Beantwortung Aufschluss über die Fähigkeiten einer Testperson gibt. Es wird unterschieden zwischen der „Klassischen Testtheorie" und der sogenannten „Probabilistischen Testtheorie" oder „Item Response Theory". Ich werde auf beide Theorien später genauer eingehen.
Die klassische Testtheorie geht davon aus, dass die Testergebnisse, die beobachtbar sind, durch die wahren Fähigkeiten der Testpersonen und den Messfehlern zustandekommen.
Bei der probabilistischen Testtheorie wird hingegen die Wahrscheinlichkeit betrachtet, mit der ein Proband mit bestimmten Fähigkeiten eine Aufgabe mit bestimmten Schwierigkeitsgrad, löst.

1.2 Faktorenanalyse und dazugehörige Modelle

Charles Spearman gilt als Erfinder der Faktorenanalyse. Er entwickelte ein Modell mit einem Faktor, der die Intelligenz darstellen sollte. Später wurden dann Modelle entwickelt, bei denen mehrere Faktoren dazu führen, dass eine Versuchsperson bei einem Test so oder so antwortet. Anhand der Korrelationen der manifesten Variablen (z.B. Antworten auf einen Test) werden die Faktoren ermittelt, die die Varianzen der Variablen erklären sollen. Die Faktoren werden z.B. mit Maximum-Likelihood geschätzt. Mit Hilfe von Rotationsverfahren werden die Faktoren iterativ den Daten angepasst, so dass die Faktoren übrigbleiben, die die Varianz am besten erklären. Also die Faktoren mit den größten sogenannten „Faktorladungen". Dazu dient z.B. das Varimax-Kriterium von Kaiser. Eine Methode von Hand ist von L.L. Thurstone (1935) entwickelt worden und ist vorallem für korrelierende Faktoren gut geeignet.
Bei der konfirmatorischen Faktorenanalyse werden zunächst die Werte einiger Faktorladungen festgelegt und dann geschaut, wie gut diese Faktoren zu den Daten passen. Das wird mit sogenannten „Strukturgleichungsmodellen" erreicht Ein wichtiges Modell bei dieser Herangehensweise ist das LISREL-Modell von Jöreskog (1973).

1.3 Multidimensionale Skalierung

Um die Distanzen zwischen psychologischen Eigenschaften untersuchen zu können benötigt man multidimensionale Skalierungen. Young und Householder entwickleten 1938 eine Methode um Punkte einer Matrix zu erhalten, die diese Distanzen enthält. Das Konzept einer beliebigen monotonen Beziehung zwischen der beobachteten Ungleichheit und den wahren Abständen wurde von Shepard 1962 vorgeschlagen und Kruskal entwickelte 1964 dazu einen Algorithmus für nichtmetrische multidimensionale Skalierung. Guttman (1968) wollte eine Rangfolge für die monotonen Beziehungen aufstellen. Carroll und Chang (1970) schlugen eine Lösung vor, bei der man mehrere Matritzen mit Ungleichheiten gleichzeitig skalieren kann, um Unterschiede zwischen Personen festzustellen. All diese Methoden beinhalten keinerlei Verteilungsannahmen.
Takana entwickelte Modelle, die mitberücksichtigen aus welchen Daten die Ungleichheiten kommen. Sie beinhalten Verteilungsannahmen und Parameterschätzer.

2 Theorie psychologischer Tests

2.1 Die Klassische Test-Theorie

Die nachfolgenden Ausführungen sind angelehnt an *Moosbrugger (2001)* und an*Bühner (2004 und 2010)* Die klassische Testtheorie wird deßhalb so bezeichnet, weil sie die grundlegendste und erste Testtheorie ist. Sie ist eine reine Messfehlertheorie. Ihre Prinzipien wurden von Gulliksen (1950) und von Lord&Novick (1968) entwickelt. Die hauptsächliche Gleichung in der klassischen Testtheorie ist:

$$x_{vi} = \tau_{vi} + \epsilon_{vi}$$

wobei x die einzige beobachtbare Größe, nämlich der beobachtete Wert ist. τ beschreibt den wahren Wert und ist wie der Fehler ϵ, der die unsystematischen Störeinflüsse beschreibt, latent. Die klassische Testtheorie ist also nur für mindestens intervallskalierte Daten sinnvoll. i ist hier und auch im Folgenden der Itemindex und v der Personenindex.

2.1.1 Annahmen der klassischen Testtheorie

Es gelten sechs Axiome, die als wesentliche Annahmen über den wahren Wert und den Messfehler dienen:

a)
$$x_{vi} = \tau_{vi} + \epsilon_{vi} \tag{1}$$
Jeder beobachtete Wert einer Person für ein Item ist zusammengesetzt aus dem wahren Wert und dem Messfehler.

b)
$$\tau_{vi} = \mathbb{E}(x_{vi}) \tag{2}$$
wobei τ_{vi} für den wahren Wert einer Person steht und x_{vi} für den beobachteten Wert.
daraus ergibt sich, dass der Erwartungswert der Messfehler Null ist.

$$\mathbb{E}(\epsilon_{vi}) = 0$$

c)
$$Corr(\tau_{vi}, \epsilon_{vi}) = 0 \tag{3}$$
Da der wahre Wert eine Konstante ist ($\mathbb{V} = 0$), kann er nicht mit den variierenden Messfehlern korrelieren. Der wahre Wert und der Messfehler hängen also nicht voneinander ab.

d)
$$Corr(\epsilon_{vi}, \epsilon_{vj}) = 0 \tag{4}$$

Die Testitems sind unabhängig voneinander. Die Bearbeitung von einem Item darf also nicht von dem Erfolg eines anderen Items beeinflusst werden. So darf z.B. nicht die Lösung einer Aufgabe notwendig für die Lösung einer anderen Aufgabe sein, sonst wäre diese Annahme verletzt.

e)
$$Corr(\epsilon_{vi}, \epsilon_{wi}) = 0 \tag{5}$$
Die Bearbeitung darf auch nicht von einer anderen Person abhängen. Z.B. wäre diese Annahme verletzt, wenn eine Person von einer anderen abschreibt.
Die Axiome d) und e) bezeichnet man auch als "lokale stochastische Unabhängigkeit".

f) da es keinen systematischen Fehler gibt, gilt: Bias = 0

Allerdings sind diese Annahmen etwas kritisch, da weder ϵ_{vi} noch τ_{vi} beobachtbar ist. Somit ist es schwierig über sie Aussagen zu treffen. Im Weiteren wird davon ausgegangen, dass die Axiome gelten.

2.1.2 Die Bestimmung des wahren Wertes

Um den wahren Wert zu bestimmen, müssen an einer Person verschiedene Items getestet werden, die allerdings alle denselben Sachverhalt messen. Das ist nicht ganz einfach, da die Fehler nach Axiom d) unabhängig sein müssen. Das gleiche Item immer wieder zu verwenden wäre also nicht möglich, da die Unabhängigkeit z.B. durch Erinnerungsleistungen verletzt wäre.
Wenn es gelungen ist solche Items zu konstruieren, werden zur Bestimmung des wahren Wertes die Testleistungen addiert und dann der Erwartungswert berechnet:

$$x_v = \sum_{i=1}^{m} x_{vi}$$

$$\begin{aligned} \mathbb{E}(x_v) &= \mathbb{E}(\sum_{i=1}^{m} x_{vi}) \\ &= \sum_{i=1}^{m} \mathbb{E}(x_{vi}) \\ &\overset{(2)}{=} \sum_{i=1}^{m} \tau_{vi} \\ &= \tau_v \end{aligned} \tag{6}$$

Der Schätzer $x_v = \hat{\tau}_v$ ist also erwartungstreu für τ_v Die Formulierung "wahrer Wert“ ist etwas kritisch, da er eigentlich nur den Erwartungswert der beobachtbaren Testergebnisse darstellt. „Erwarteter Wert“ wäre vielleicht eine bessere Bezeichnung.

2.1.3 Reliabilität

Die Reliabilität gibt an wie genau „ein Test die Eigenschaften misst, die er tatsächlich misst." *(Kranz, 1979, S.4)*. Eine Person sollte verschieden Fragen, die aber das gleiche messen, immer „in der gleichen Art und Weise beantworten" *(Bühner, 2004, S.24)*. Im Unterschied zur Validität eines Tests, die angibt wie genau „ein Test das misst, was er messen soll." *(Kranz, 1979, S.4)*. Zur Berechnung der Reliabilität bedient man sich häufig parallelen Test. Parallele Tests messen das gleiche Konstrukt. Ihre Korrelation gibt Aufschluss über die Genauigkeit der Tests. Zwei solcher Tests p und q „messen die gleichen wahren Werte" *(Kranz, 1979, S. 85)*:

$$\tau_{p_v} = \tau_{q_v} \tag{7}$$

wobei τ_{p_v} der wahre Wert einer Person v in Test p und τ_{q_v} der wahre Wert einer Person v in Test q ist. daraus ergibt sich:

$$\mathbb{V}(\tau_p) = \mathbb{V}(\tau_q)$$

und:

$$\overline{\tau_p} = \overline{\tau_q}$$

Weiterhin sind die Standardmessfehler zweier paralleler Tests gleich:

$$SD(\epsilon_p) = SD(\epsilon_q) \tag{8}$$

(Kranz, 1979) Die Reliabilität gibt das Verhältnis der Varianz des wahren Werts zur Varianz des beobachteten Werts wider.

$$Rel = \frac{\mathbb{V}(\tau)}{\mathbb{V}(x)} \tag{9}$$

Es handelt sich also genau genommen um einen Reliabilitätskoeffizienten, mit Wertebereich zwischen 0 und 1. Je näher er an 1 ist, desto genauer ist der Test.

$$\mathbb{V}(x) = \mathbb{V}(\tau) \longrightarrow Rel = 1$$

$$\mathbb{V}(x) = \mathbb{V}(\epsilon) \longrightarrow Rel = 0$$

Es gibt vier Methoden zur Reliabilitätsschätzung:

- Paralleltest-Reliabilität
- Retest-Reliabilität
- Split-Half-Reliabilität
- Interne Konsistenz

2.1.4 Konfidenzintervall für τ_v

Um eine möglischst genaue Aussage über τ_v treffen zu können, muss man zusätzlich zu dem oben berechneten Schätzer $\hat{\tau}_v$ ein Konfidenzintervall berechnen. Dazu benötigen wir zunächst den Standardmessfehler $SD(\epsilon)$.
Gleichung (9) lässt sich wie folgt umformen:

$$\mathbb{V}(x) = Rel \cdot \mathbb{V}(x) + (1 - Rel) \cdot \mathbb{V}(x)$$

$$\Rightarrow \mathbb{V}(\epsilon) = \mathbb{V}(x) \cdot (1 - Rel)$$

Den Standardmessfehler kann man damit leicht berechnen:

$$SD(\epsilon) = SD(x) \cdot \sqrt{1 - Rel} \tag{10}$$

folglich wird $SD(\epsilon)$ mit zunehmender Reliabilität kleiner.

Wir nehmen nun an, dass die Fehler normalverteilt sind und eine ausreichend große Stichprobe zu Verfügung steht. Dann sieht das Konfidenzintervall folgendermaßen aus:

$$KI = \left[\hat{\tau}_v \pm z_{\alpha/2} \cdot SD(\epsilon)\right] \tag{11}$$

2.1.5 Kritik an der klassischen Testtheorie

- Wie bereits erwhnt, ist weder τ_{vi} noch ϵ_{vi} beobachtbar. Also können wir auch nicht mit Sicherheit sagen, dass die Axiome gelten. Auf diesen Annahme baut jedoch die Theorie auf.
- Einige Kennwerte wie z.B. die Schwierigkeit der Items ist stichprobenabhängig. Eine Stichprobe, die viele sehr fähige Probanden enthält wird die Aufgaben leichter bewältigen als eine Stichprobe mit weniger begabten Probanden und daher lassen sich die Ergebnisse schlecht verallgemeinern. Die Schwierigkeit einer Aufgabe würde bei zwei Gruppen mit unterschiedlicher mittlerer Fähigkeit unterschiedlich eingestuft werden.
- Die Klassische Testtheorie berücksichtigt nur die unsystematischen Fehler. Systematische Fehler, die z.B. durch Lerneffekte auftreten, werden nicht in das Modell aufgenommen.

2.2 Die Probabilistische Test-Theorie

Bei der probabilistischen Testtheorie wird die Schwierigkeit einer Aufgabe, wie wir später sehen werden, mit einer eigenen Variablen angegeben und nicht wie bei der klassischen Testtheorie als „Anteil richtiger Antworten auf das jeweilige Item“ *(Filipp, 1993, S.45)*. Außerdem werden die Fähigkeiten der Probanden auch mitberücksichtigt. Zusätzlich wird noch der Zufall mit einbezogen, dass dieselbe Person nicht immer gleich auf die Fragen antwortet. Es wird hier die Wahrscheinlichkeit betrachtet ein bestimmtes Item zu lösen.
Man geht davon aus, dass die manifeste Variable „Antworten auf verschiedene Items“ von der latenten Variablen „Fähigkeit“ abhängt. Es ist wichtig, dass die Antworten nur durch die Fähigkeiten „systematisch beeinflusst werden“ *(Moosbrugger, 2008)*

2.2.1 Das Rasch-Modell

Die folgenden Ausführungen basieren hauptsächlich auf *Strobl, 2010*. Die Daten für die Grafiken stammen von*http://www.sozialwissenschaftliche-forschungsmethoden.de*, es handelt sich um die*Beispiel-Rasch-Daten*
Um einen Test zu konstruieren, ist es wichtig, dass die Aufgaben so gestellt sind, dass sie nur die interessierende Eigenschaft messen und sich keine weiteren Fähigkeiten der Probanden auf das Ergebnis auswirken. Zur Überprüfung, ob die Items dieser Anforderung gerecht werden, dient das Rasch-Modell. Die Antworten der Items nehmen die Werte 0 und 1 an, sind also dichotom. Um ein lineares Modell mit einer binären Zielgröße aufzustellen, bedient man sich dem Logit-Modell mit der Funktion:

$$\frac{exp(x)}{1 + exp(x)}$$

Der Wertebereich der logistischen Funktion liegt zwischen 0 und 1 und sie bildet eine Wahrscheinlichkeit ab. Eine weitere Eigenschaft des Logit-Modells ist, dass seine Funktion mit zunehmendem x steigt. Wie wir gleich sehen werden, benötigen wir diese Eigenschaften für das Rasch-Modell. Die Gleichung des Rasch-Modells sieht folgendermaßen aus:

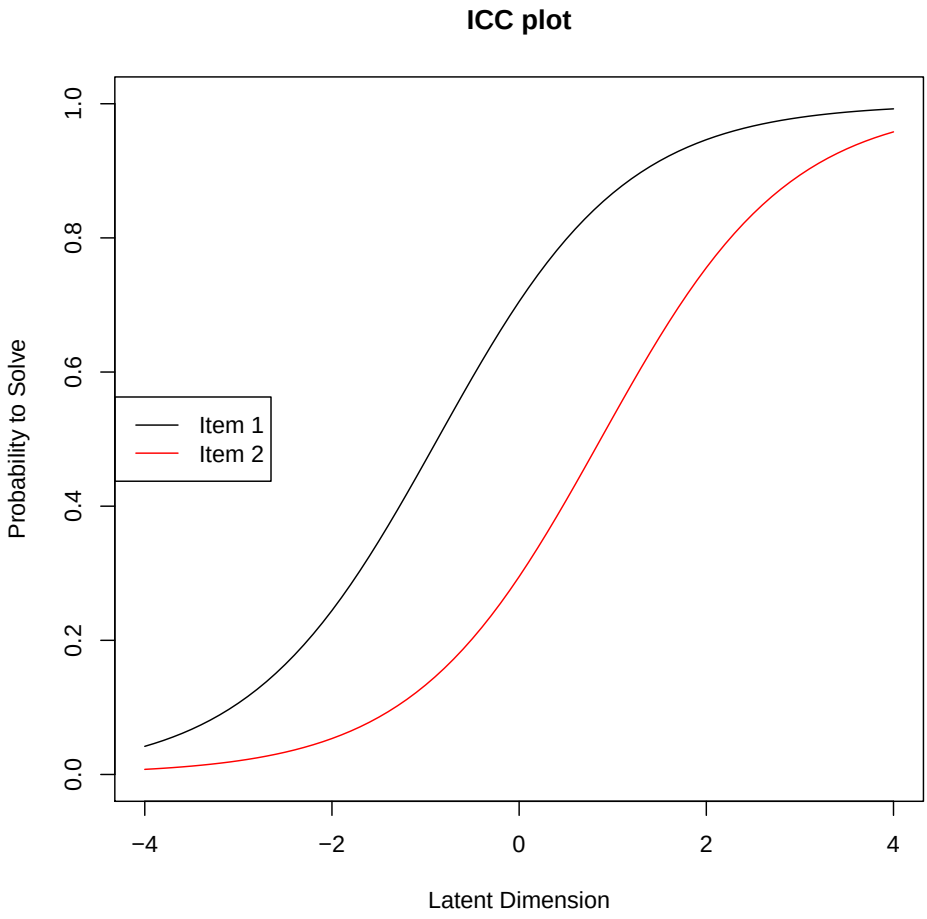
$$\mathbb{P}(U_{ij} = 1|\vartheta_i, \beta_j) = \frac{exp(\vartheta_i - \beta_j)}{1 + exp(\vartheta_i - \beta_j)} \tag{12}$$

Die Gleichung beschreibt die Wahrscheinlichkeit eine Aufgabe richtig zu lösen, wobei ϑ_i die Fähigkeit der Person i und β_j die Schwierigkeit einer Aufgabe j darstellt.
Wenn man offen lassen will, ob die Aufgabe gelöst wird oder nicht, gilt die Formel:

$$\mathbb{P}(U_{ij} = u_{ij}|\vartheta_i, \beta_j) = \frac{exp(u_{ij} \cdot (\vartheta_i - \beta_j))}{1 + exp(\vartheta_i - \beta_j)} \tag{13}$$

wobei u_{ij} beschreibt, ob eine Person i eine Aufgabe j gelöst hat oder nicht.
$\vartheta_i - \beta_j$ ist also x aus dem Logit-Modell. Ist jetzt $\vartheta_i > \beta_j$ und somit $x > 0$, steigt

die Funktion und somit auch die Wahrscheinlichkeit, die Aufgabe zu lösen. Im umgekehrten Fall, $\vartheta_i < \beta_j$, sinkt die Lösungswahrscheinlichkeit. So wie gefordert bezieht also das Rasch-Modell die Fähigkeit einer Person und die Schwierigkeit der Aufgabe in die Wahrscheinlichkeit mit ein, eine Aufgabe richtig zu lösen.
Um die Lösungswahrscheinlichkeit einer Aufgabe darzustellen, verwendet man sogenannte „Item Cahrakteristic Curves"(kurz: ICC). Die Form dieser ICCs hängt von der Schwierigkeit der Aufgabe ab. Sie stellen die Wahrscheinlichkeit eine Aufgabe zu lösen in Abhängigkeit der Fähigkeit einer Person dar.
Die Form der Funktion ist auch bei mehreren Aufgaben in einem Test gleich. Lediglich der Schweregrad kann variieren. Wenn bei Aufgaben noch zusätzliche Faktoren in die Form der Funktion mit eingehen, entsprechen sie nicht dem Rasch-Modell und werden aus dem Test eliminiert. Da die Form also immer gleich sein muss, ist auch die „Steigung im mittleren Bereich, die als Trennschärfe bezeichnet wird" *(Strobl, 2010, S.11)*, immer gleich. Die Trennschärfe gibt an, wie „genau man mit einer Aufgabe zwischen Personen mit unterschiedlichen Fähigkeiten unterscheiden kann" *(*Strobl, 2010, S.11). (s. Grafik unten)



Die Aufgaben in einem Test müssen bestimmte Bedingungen erfüllen, die im Rasch-Modell gelten müssen. Eine dieser Bedingungen ist die lokale stochastische Unabhängigkeit, die auch schon bei der klassischen Testtheorie gelten musste. Die Aufgaben und die Personen, die diese Aufgaben lösen, dürfen also wieder nicht zusammenhängen. Wenn man von lokaler stochastischen Unabhängigkeit ausgeht, kann man die Wahrscheinlichkeit für mehrere Personen (1,....n), verschiedene Aufgaben (1,....m) zu lösen folgendermaßen berechnen:

$$\begin{aligned} \mathbb{P}(U = u|\vartheta, \beta) &= \prod_{i=1}^{n} \prod_{j=1}^{m} \mathbb{P}(U_{ij} = u_{ij}|\vartheta_i, \beta_j) \\ &= \frac{\exp(\sum_{i=1}^{n} r_i \cdot \vartheta_i - \sum_{j=1}^{m} s_j \cdot \beta_j)}{\prod_{i=1}^{n} \prod_{j=1}^{m} (1 + \exp(\vartheta_i - \beta_j))} \end{aligned} \tag{14}$$

r_i ist hier die Summe der gelösten Aufgaben einer Person i und s_j die Summe der Personen, die eine Aufgabe j gelöst haben. r_i und s_j sind jeweils suffiziente Statistiken für den Personen- bzw. den Aufgabenparameter. Aus dieser Gleichung könnte sich z.B. folgende Beispielmatrix ergeben:

	Aufgabe				
Person	1	2	3	4	r_i
1	1	1	0	0	2
2	0	1	1	1	3
3	1	0	0	1	2
s_j	2	2	1	2	

Eine weitere Bedingung für die Aufgaben in einem Test ist die sogenannte „Spezifische Objektivität“ *(Strobl, 2010)*. Diese besagt, dass bei einem Vergleich von zwei Personen nicht die eine Person bei einer Aufgabe als „Sieger“ hervorgeht und bei einer anderen als "Verlierer". Der Vergleich der Fähigkeiten muss bei allen Aufgaben gleich ausgehen. Das ist gewährleistet, wenn die Funktionen der Lösungswahrscheinlichkeiten parallel verlaufen, wie in Bild 2 dargestellt. In Formeln ausgedrückt sieht das bei zwei Personen a und b und zwei Aufgaben j und j' folgendermaßen aus:

$$\frac{\frac{\mathbb{P}(u_{aj}=1|\vartheta_a,\beta_j)}{1-\mathbb{P}(u_{aj}=1|\vartheta_a,\beta_j)}}{\frac{\mathbb{P}(u_{bj}=1|\vartheta_b,\beta_j)}{1-\mathbb{P}(u_{bj}=1|\vartheta_b,\beta_j)}} = \exp(\vartheta_a - \vartheta_b) \tag{15}$$

$$= \frac{\frac{\mathbb{P}(u_{aj'}=1|\vartheta_a,\beta_{j'})}{1-\mathbb{P}(u_{aj'}=1|\vartheta_a,\beta_{j'})}}{\frac{\mathbb{P}(u_{bj'}=1|\vartheta_b,\beta_{j'})}{1-\mathbb{P}(u_{bj'}=1|\vartheta_b,\beta_{j'})}}$$

Die Odds Ratio hängt also nicht von der Aufgabenschwierigkeit ab, sondern nur noch von den Fähigkeiten der Personen.

Eine Möglichkeit, die Personen- und Aufgabenparameter aus einer Stichprobe zu schätzen, ist die gemeinsame Maximum-Likelihood-Schätzung:

$$L_u(\vartheta,\beta) = \prod_{1=1}^{n}\prod_{j=1}^{m}\frac{\exp(u_{ij}\cdot(\vartheta_i-\beta_j))}{1+\exp(\vartheta_i-\beta_j)} \tag{16}$$

$$= \frac{\exp(\sum_{i=1}^{n} r_i\cdot\vartheta_i - \sum_{j=1}^{m} s_j\cdot\beta_j)}{\prod_{i=1}^{n}\prod_{j=1}^{m}(1+\exp(\vartheta_i-\beta_j))}$$

Die damit berechneten Schätzer sind allerding nicht konsistent, da mit steigender Stichprobengröße auch die Anzahl der Parameter steigt, die man schätzen soll. Man kann also nicht „eine gleichbleibende Anzahl von Parametern mit einer größer werdenden Stichprobe schätzen” *(Strobl, 2010, S. 29)*. Bei konsistenten Schätzern wird die Varianz geringer mit steigender Stichprobengröße. Das ist hier nicht der Fall.

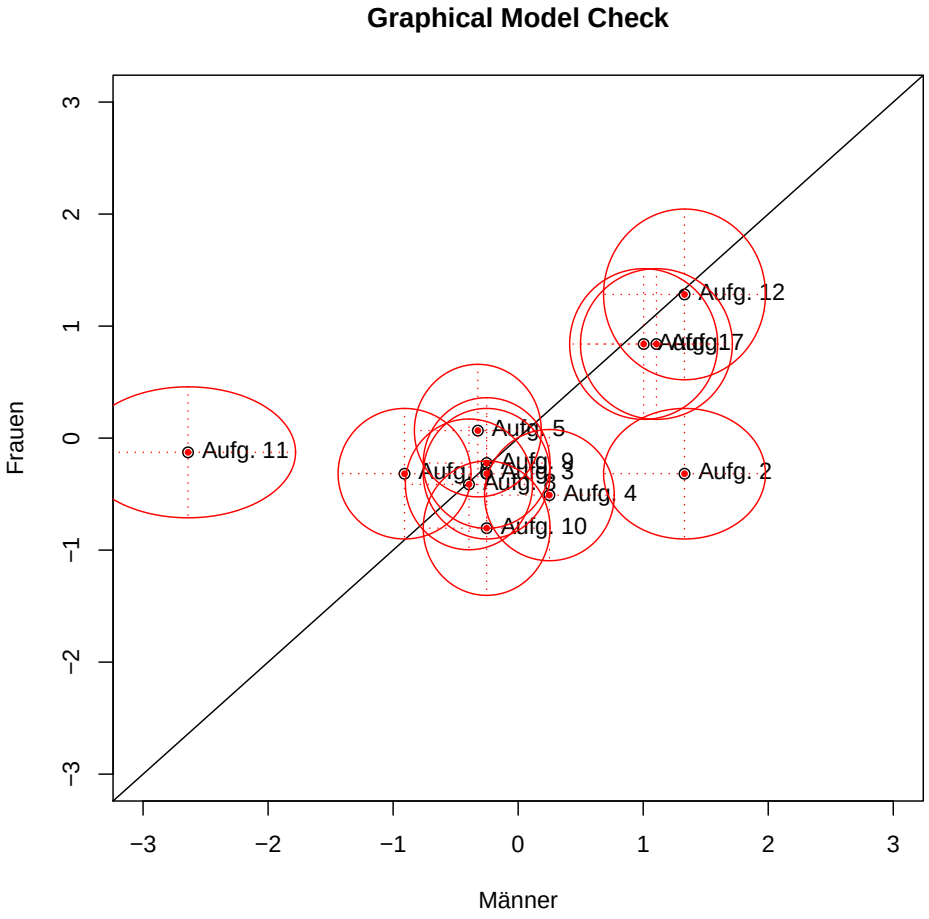
Eine weitere Möglichkeit ist die bedingte Maximum-Likelihood-Schätzung, bei der im ersten Schritt der Aufgabenparameter geschätzt wird anhand der bedingten Likelihood:

$$\begin{aligned} h(u|r,\beta) &= \prod_{i=1}^{n} h(u_i|r_i,\beta) \\ &= \prod_{i=1}^{n} \frac{exp(-\sum_{j=1}^{m} u_{ij} \cdot \beta_j)}{\gamma_{r_i}(\beta)} \\ &= \frac{\exp(-\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ij} \cdot \beta_j)}{\prod_{i=1}^{n} \gamma_{r_i}(\beta)} \\ &= \frac{\exp(-\sum_{j=1}^{m} s_j \cdot \beta_j)}{\prod_{i=1}^{n} \gamma_{r_i}(\beta)} \end{aligned} \tag{17}$$

mit $\gamma_{r_i}(\beta) = \sum_{\sum_j u_{ij}=r_i} exp(-\sum_{j=1}^{m} u_{ij} \cdot \beta_j)$

Diese Funktion hängt nicht mehr vom Personenparameter ϑ_i ab und die Aufgabenparameter können mittels iterativer Maximum-Likelihood-Schätzung geschätzt werden. Anhand der geschätzten Aufgabenparameter werden dann die Personenparameter geschätzt.
Auch bei der sogenannten "Marginalen Maximum-Likelihood-Schätzung"werden zunächst die Aufgabenparameter geschätzt. Dazu wird die Randdichte der Personenparameter mit der Likelihood-Funktion multipliziert und dann über ϑ_i integriert. Dadurch erhält man eine Funktion, die "nur noch die Aufgabenparameter als unbekannte Größen enthält" *(Strobl, 2010, S.34)* und man kann den Maximum-Likelihood-Schätzer ausrechnen. Für die Berechnung der Randdichte nimmt man die Standard-Normalverteilung für die Personen-Parameter an. Im zweiten Schritt werden dann wieder die Personenparameter geschätzt.

Um nun herauszufinden, ob die Aufgaben den Anforderungen des Rasch-Modells genügen, kann man verschieden Tests durchführen. Einer dieser Tests ist der sogenannte "graphische Modelltest". Bei dem geschaut wird, ob bei einzelne Aufgaben die geschätzten Parametern bei zwei Personengruppen systematisch voneinander abweichen. Dazu werden zwei Personengruppen gegeneinander in einem Koordinatensystem aufgetragen. Die Parameterschätzer sollten dann ungefähr auf der Winkelhalbierenden liegen. Es werden nicht nur die Schätzer sondern auch ihre "Konfidenz-Regionen" *(Strobl, 2010, S.41)* abgebildet. Diejenigen Aufgaben, deren Konfidenz-Regionen die Winkelhalbierende schneiden, folgen den Anforderungen des Rasch-Modells und können im Test bleiben. Die anderen müssen herausgenommen werden, weil sie von den Personen abhängig sind. (s. Grafik unten)



Das Problem des graphischen Modelltests ist, dass er immer nur zwei Personengruppen gleichzeitig untersuchen kann. Der Likelihood-Quotienten-Test kann hier Abhilfe schaffen.
Der Likelihood-Quotient vergleicht den Schätzer des Aufgabenparameters, wenn er gemeinsam über alle Personen geschätzt wird und den Schätzern, die sich für einzelne Personengruppen $k = 1, ..., K$ ergeben.

$$LQ = \frac{L_u(r, \hat{\beta})}{\prod_{k=1}^{K} L_{u_k}(r_k, \hat{\beta}_k)}$$

Die dazugehörige Teststatistik ist:

$$T = -2lnLQ$$

die χ^2-verteilt ist. Wenn das Rasch-Modell gilt, dann ist $LQ = 1$ und somit $T = 0$. Wenn das Rasch-Modell verletzt ist, ist $T > 0$. Demnach sprechen große Werte von T gegen das Rasch-Modell.

Literatur:

Bühner, Markus: Einführung in die Test- und Fragebogenkonstruktion, 1. Auflage, Pearson Studium: 2004
Bühner, Markus: Einführung in die Test- und Fragebogenkonstruktion, 1. Auflage, Pearson Studium: 2010
Filipp, Gernot: Probabilistische Testmodelle in der Persönlichkeitsdiagnostik, Peter Lang GmbH: 1993
Kranz, Helger T.: Einführung in die klassische Testtheorie, 1. Auflage, Fachbuchhandlung für Psychologie GmbH: 1979
Moosbrugger, Helfried und Kelava, Augustin (Hrsg.): Testtheorie und Fragebogenkonstruktion, Springer: 2007
Strobl, Carolin: Das Rasch-Modell, 1. Auflage, Rainer Hampp: 2010