5. Die probabilistische Testtheorie (PTT)

Aus der Kritik der klassischen Testtheorie wurden eine Reihe "moderner" Testtheorien entwickelt. Da hier Aussagen über Auftretenswahrscheinlichkeiten von manifestem, beobachtbarem Verhalten gemacht wird, spricht man auch von probabilistischen oder stochastischen Modellen. Man geht davon aus, dass eine Person eine Aufgabe nur mit einer bestimmten Wahrscheinlichkeit lösen wird.
Messen wird hier als Vorgang des Vergleichens aufgefasst. Verglichen werden 2 Parameter (auch Determinanten genannt), nämlich die Fähigkeit der Person und die Schwierigkeit eines Items. Fähigkeit (z.B. Intelligenz, Einstellung, Persönlichkeit) und Schwierigkeit werden als latente, nicht direkt beobachtbare Variablen angesehen, über die man aufgrund der manifesten, direkt beobachtbaren Reaktion auf das Testitem schliessen kann.
Formal wird dies wie folgt ausgedrückt:

$$P(A_{vi} = 1) = f(\xi_v, \sigma_i)$$

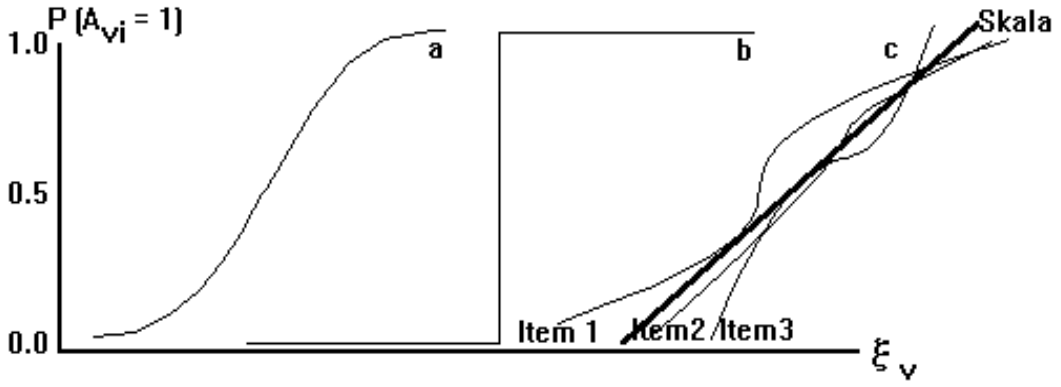
ξ_v ... Fähigkeit der Person
σ_i ... Schwierigkeit des Items i

Die Reaktionswahrscheinlichkeit ist eine Funktion aus Fähigkeit der Person und Itemschwierigkeit. Beobachtbar ist die Wahrscheinlichkeit, ein Item i zu lösen. Latente, nicht beobachtbare Variablen sind die Fähigkeit und die Schwierigkeit.

5.1. Die Itemcharakteristik als zentrales Konzept probabilistischer Testmodelle

Betrachtet man die Reaktionswahrscheinlichkeit hinsichtlich eines Items i, kann man die beobachtbare Variable , nämlich die Beantwortungswahrscheinlichkeit gegen die latente Variable, nämlich die Fähigkeit der Person auftragen. Die Funktion f legt die Beziehung zwischen beiden Dimensionen fest und wird Itemcharakteristik oder Itemkennlinie (ICC) genannt.

Abbildung 11: Itemcharakteristiken für verschiedene Testmodelle



Legende: Dargestellt ist die Beantwortungswahrscheinlichkeit P (A_{vi} = 1) in Abhängigkeit von der Personenfähigkeit ξ_v für verschiedene Testmodelle.

[a] Rasch-Modell (1960)

[b] Guttman-Skalierung (1950)

[c] Klassische Testtheorie

Für jedes Messmodell muss man definieren durch welche Funktion die Beantwortungswahrscheinlichkeit mit der Fähigkeit der Person zusammenhängt. Innerhalb eines Testmodells wird dieselbe Itemcharakteristik für alle Items angenommen.

Fall a)

a stellt die ICC des Rasch-Modells (1960) dar. Die Wachstumskurve ist eine logistische Funktion. Mit zunehmender Fähigkeit steigt die Wahrscheinlichkeit für eine richtige Antwort P(A_{vi} = 1) monoton und stetig an. Für die Extrembereiche der Fähigkeit nähert sich die Funktion den Werten P=1 und P=0 asymptotisch an. Es ist eine echte Wahrscheinlichkeitsfunktion.

Fall b)

b zeigt das deterministische Modell von Guttman (1950). Die Reaktionswahrscheinlichkeit kann nur die Werte P=0 oder P=1 annehmen. Unterscheiden sich die Items hinsichtlich ihrer Schwierigkeit, drückt sich dies in einer Verschiebung der Sprungstelle aus. Eine Person, die ein bestimmtes Item gelöst hat, muss auch alle leichteren gelöst haben. Wenn die Person ein Item nicht gelöst hat, darf sie auch kein schwereres lösen (= Homogenität). Man kann dies auch in einer Datenmatrix darstellen:

Tabelle 6: Modellgeltungstest bei Guttman-Skalierung

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

Anmerkung: Die Items 1 bis 5 wurden von 4 Personen beantwortet. 1 bedeutet, dass das Item gelöst wurde, 0 bedeutet nicht gelöst.

Wenn man eine Matrix wie in Tabelle 6 herstellen kann (durch Vertauschen von Spalten und Zeilen), hat man den Modellgeltungstest durchgeführt. In der Praxis gibt es kaum Tests die diese Forderung erfüllen. Bsp. für Items: Item1: "Gleiche Berufschancen für Farbige" bis Item 4: "Ich würde einen Farbigen heiraten".

Fall c)
c beschreibt die Itemkennlinien für die klassische Testtheorie. Sie verwendet monotone Modelle mit nicht spezifischer Verteilungsannahme. Die Skala oder der Itemsummenwert soll aber dann einen linearen Zusammenhang zur latenten Eigenschaft (Fähigkeit) zeigen.

In der klassischen Testtheorie wurde die Schwierigkeit eines Items i definiert als Anzahl richtiger Lösungen auf dem Item durch die Gesamtzahl der Antworten auf dem Item. Die Schwierigkeit hängt dadurch stark von der Personenstichprobe ab. Im Rasch-Modell und bei der Guttman-Skalierung wird die Schwierigkeit eines Items im Vergleich zur Schwierigkeit eines oder mehrerer anderer Items bestimmt (Parallelverschiebung der ICC überführt leichtes in schwieriges Item und umgekehrt).

5.2. Das Rasch-Modell

Das Rasch-Modell beschreibt eine logistische Funktion (eine e-Funktion) aus Personenfähigkeit und Itemschwierigkeit. Die Modellgleichung lautet (Rasch, 1960):

$$P(A_{vi} = 1) = \frac{e^{x_v - s_i}}{1 + e^{x_v - s_i}}$$

e ... Eulersche Zahl (2,72)
ξ_v ... Fähigkeit der Person v
σ_i ... Schwierigkeit Item i

Die Gleichung beschreibt die Funktion der Itemkennlinie a in Abbildung 11.
Mit dieser Gleichung und dem Postulat der lokalen stochastischen Unabhängigkeit der Reaktionen ist das Modell von Rasch festgelegt.

5.2.1. Die Annahme der lokalen stochastischen Unabhängigkeit

Die Lösungswahrscheinlichkeit einer Person eines Items 2 hängt mit deren Fähigkeit zusammen und nicht mit der richtigen oder falschen Beantwortung eines vorher bearbeiteten Items 1. Dies bedeutet, dass die Beobachtungen unabhängig voneinander gemacht werden, also sich nicht gegenseitig beeinflussen. Dies ist wichtig für die Berechnung der Wahrscheinlichkeit.

5.2.2. Der Testscore als erschöpfende Statistik

Die Skalen der klassische Testtheorie sind üblicherweise Summen aus Einzelitems. Dabei spielt es natürlich keine Rolle in welcher Sequenz die Items beantwortet wurden (Bsp.: Lösungsvektor 1: (1 0 0 1 1) = Summenwert 3 = Lösungsvektor 2: (1 1 0 1 0)). Im Modell der klassischen Testtheorie ist dieses Prinzip implizit enthalten. Im Rasch-Modell geht das Prinzip explizit als Annahme mit ein. Die Anzahl gelöster Items ist eine erschöpfende Statistik für die Fähigkeitsparameter der Person. Praktisch hat dies zu Folge, dass für die Parameterschätzung (Fähigkeit und Schwierigkeit) nur die Randvektoren der Datenmatrix (Personen x Items) benötigt werden (Spaltensummen und Zeilensummen).

5.2.3. Spezifische Objektivität und Stichprobenunabhängigkeit

Die Parameterschätzung (Fähigkeit und Schwierigkeit) ist unabhängig von der Itemsstichprobe und der Personenstichprobe (Wechselseitige Stichprobenunabhängigkeit). Die Genauigkeit der Schätzung hängt allerdings von der Stichprobengrösse ab.
Aus dieser Stichprobenunabhängigkeit folgt, dass der Vergleich zweier Personen dann als spezifisch objektiv gilt, wenn das Ergebnis des Vergleichs nicht vom Instrument abhängig ist, mit dessen Hilfe er durchgeführt wurde. Der Vergleich ist unabhängig von der Itemauswahl oder der Personenauswahl.
Bsp.: Die Gewichtsbestimmung einer Kugel soll unabhängig davon sein, ob eine Federwaage oder eine Balkenwaage verwendet wird. Zudem sollte das Ergebnis davon unabhängig sein, ob noch andere Kugeln gemessen werden.

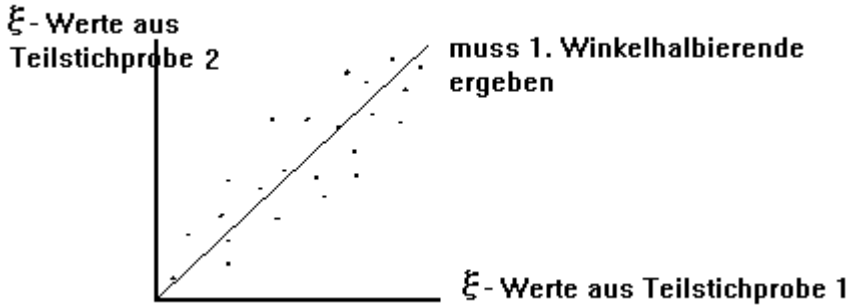
5.2.4. Parameterschätzung

Die Parameter der logistischen Funktion des Rasch-Modells werden durch eine bedingte Maximum-Likelihood-Schätzung ermittelt. Der Schätzvorgang erfolgt iterativ (immer genauere Anpassung an den Schätzwert; Bsp: Kreisbeschreibung durch Geraden) und kann praktisch nur mit einem Computer durchgeführt werden (Testat von SYSTAT). Die Schätzgenauigkeit kann wie in der klassischen Testtheorie über Konfidenzintervalle bestimmt werden, im Gegensatz zur KTT sind die Intervalle im Extrembereich grösser (Messung ist ungenauer) und sie hängen nicht von der Stichprobe der Items oder Personen ab, sondern von der Grösse der Stichprobe.

5.2.5. Modellgeltungstests

Der Vergleich von Itemparametern, die aus verschiedenen Stichprobensegmenten geschätzt wurden stellt die Kontrolle des Modells dar. Die statistische Signifikanz wird durch den Likelihoodquotiententest (Kennwerte aus Teilstichproben und Gesamtstichprobe müssen gleich sein) geprüft. Zudem ist eine graphische Kontrolle möglich:

Abbildung 12: Graphischer Modellgeltungstest des Rasch-Modells



Legende: Aus einer Teilstichprobe 1 der Items werden die Fähigkeiten der Personen bestimmt. Ergibt eine Teilstichprobe 2 der Items entsprechende Fähigkeiten für die selben Personen, so ist der Test nach dem Modell von Rasch gültig.

Die Unabhängigkeit der Parameterschätzung von den Stichproben, die im Modellgeltungstest geprüft wird, hat den grossen Vorteil, dass man individualisiert testen kann, und dass man die Probleme der Veränderungsmessung der KTT nicht hat. Im individualisierten Testen kann man durch Auswahl weniger Items gezielt auf die Fähigkeit der Person schliessen. Bei der Messung von Veränderungen sind die Tests zu Zeitpunkt 1 und Zeitpunkt 2 direkt vergleichbar, da die Parameter unabhängig von Itemauswahl und Personenauswahl geschätzt werden.

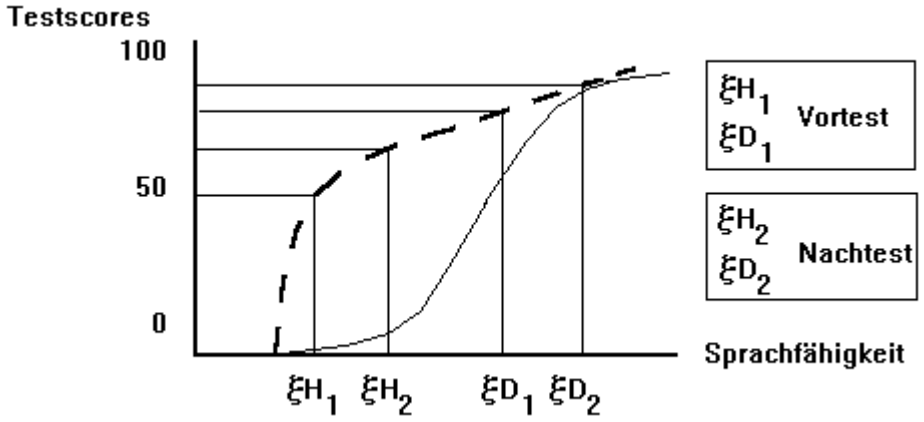
Diese beiden Vorteile der PTT werden vor allem in der Praxis relevant. Allgemein kann man allerdings feststellen, dass die PTT in der Praxis sehr wenig verwendet wird. Lippert, Schneider und Wakenhut (1977) haben die Stabilität probabilistischer Skalierungsverfahren (u.a. die Raschskalierung) anhand einer Einstellungsskala "Unpolitische Haltung" (Ellwein, Lippert & Zoll, 1975) nachzuweisen versucht. Das Freiburger Persönlichkeitsinventar wurde nach dem Raschmodell reanalysiert (vgl. Fahrenberg, Ewert & Maier, 1987). Rost und Spada (1978) zeigen eine wichtige Anwendung der probabilistischen Testtheorie zur Messung von Lerneffekten in der Pädagogik. Für den Kinder-Angst-Test (KAT) und für den Fragenbogen für Schüler (FS 5-10) hat Conrad (1976) durch Isolation von 11 aus ursprünglich 19 Items bzw. 24 aus urprünglich 38 Items schliesslich Rasch-homogene Tests entwickelt.
Das Modell ist sehr restriktiv und an hohe Voraussetzungen gebunden (z. B.: grosse Stichproben). Deshalb werden viele Items ausgesondert, die nach dem Modell der KTT noch verwendet werden würden.
Fragen der Validität werden durch das Modell nicht berührt. In der Praxis werden daher eher Tests angewandt, die ihre Validität bzgl. eines Kriteriums gezeigt haben, auch wenn sie nach dem weniger restriktiven Modell der KTT konstruiert wurden.

5.3. Uebungsaufgaben zur probabilistischen und klassischen Testtheorie

1. Die klassische und die probabilistische Testtheorie gehen von einem gemeinsamen, psychologischen Modell aus. Welches psychologische Modell ist gemeint?

2. Erkläre kurz den Unterschied zwischen klassischer und probabilistischer Testtheorie.

3. Im Rahmen der Evaluation eines Sprachförderprogrammes für milieugeschädigte Kinder wurde vor und nach der Förderphase ein Sprachtest appliziert. Wir haben die Ergebnisse für zwei Kinder, Hans (H) und Dieter (D) herausgegriffen:



Betrachte zunächst nur die dicke gestrichelte Linie. Sie gibt die Beziehung zwischen Test und Sprachfähigkeit wieder. Wenn wir diese Beziehung zugrundelegen: Wer hat mehr vom Förderprogramm profitiert?

a) auf der Fähigkeitsebene: .....
b) auf der Testscoreebene: .....

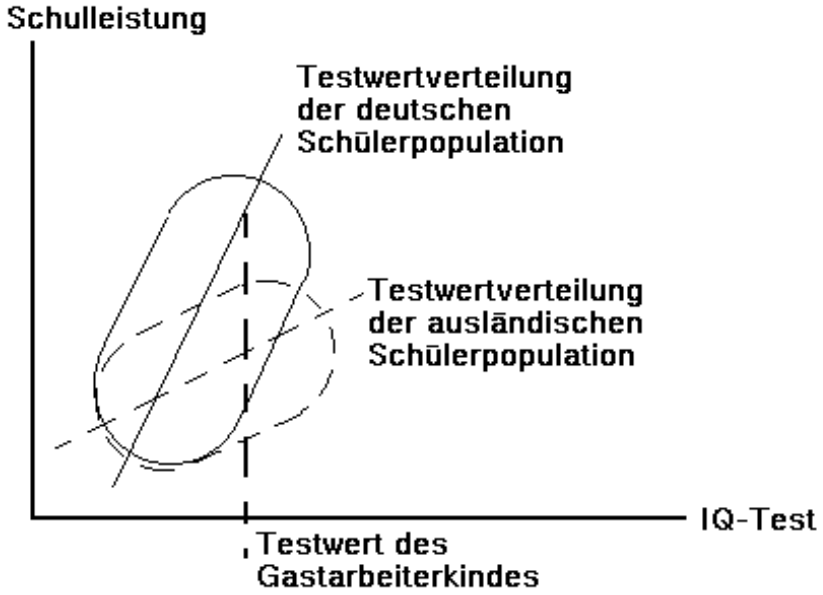
Nun zur dünnen, durchgezogenen Linie. Sie basiert auf dem gleichen Test (!), nur dass nun einige "leichte" Items durch "schwierige" ersetzt wurden. Für wen war nun das Förderungsprogramm wirksamer?

c) auf der Fähigkeitsebene: .....
d) auf der Testscoreebene: .....

e) Wenn wir das RASCH-Modell zugrundelegen: Gegen welche Voraussetzung eines wissenschaftlichen Vergleichs auf der Testscoreebene wird verstossen?

- ☐ Homoscedastizität
- ☐ Intervallskalierung
- ☐ spezifische Objektivität

4. Die Abbildung zeigt die Verteilung der Testergebnisse bei einem deutschsprachigen Intelligenztest in Relation zur Schulleistung in deutschen Schulen bezogen auf die Population deutscher Schüler (durchgezogene Linie) und bezogen auf die Population von Gastarbeiterschülern (gestrichelte Linie).



Es ist der Testwert eines Gastarbeiterkindes auf dem Intelligenztest eingetragen (dicke gestrichelte Linie). Falls man für dieses Kind die Normen des Tests für die deutsche Population verwendet, kommt man zu folgendem Resultat:

- ☐ Ueberschätzung der Schulleistung
- ☐ exakte Schätzung der Schulleistung
- ☐ Unterschätzung der Schulleistung

5. a) Zeichne ein Diagramm (x-Achse: ξ_V ; y-Achse: $P(A_{vi} = 1)$ wo die Itemcharkteristiken einer Guttman-Skala, Likert-Skala und die des Rasch-Modells eingetragen sind.

   b) Zeichne jetzt eine Versuchsperson V ein, die eine mittlere Fähigkeitsausprägung hat. Gib an wie gross die Wahrscheinlichkeit der richtigen Beantwortung bei jedem Modell in Deinem Diagramm ist.