2 Item-Response-Theorie

Die besondere Leistung der Item-Response-Theorie beruht auf einer spezifischen Bearbeitung der Relation von Person und Test. Während die klassische Testtheorie darauf abzielt, den Test an einer Eichpopulation zu einem neutralen Messinstrument der gegebenen Antworten zu machen, verlagert die Item-Response-Theorie die Funktion der Neutralität auf die reziproke Eichung von Messinstrument und Messobjekt. Das besondere Merkmal des Konzepts besteht darin, dass es die Reziprozität von Instrument und Objekt explizit in den Vordergrund rückt.

Für das Beispiel des Rasch-Modells kann man die Konstrukte Kompetenz und Schwierigkeit nur dann verknüpfen, wenn man akzeptiert, dass beide Konstrukte sich in der Form von hierarchischen Ordnungen anzeigen und auf Daten beziehen lassen. Schwierigkeiten sind danach Schwierigkeiten in dem Maße wie eine von unten nach oben *abnehmende* Zahl von Personen in der Lage ist, diese Schwierigkeiten zu bewältigen. Kompetenzen sind Kompetenzen in dem Maße wie eine von unten nach oben *zunehmende* Zahl von Aufgaben bei entsprechend kompetenten Personen ihre Lösung erfährt.

Im Folgenden stehen die formalen Grundlagen der in der Einleitung diskutierten methodologischen Aspekte der Item-Response-Theorie und der klassischen Testtheorie im Vordergrund.

Für die Item-Response-Theorie, der das Rasch-Modell angehört, basieren die formalen Grundlagen auf drei basalen Prämissen:

1) der Verwendung einer Item-Response-Funktion, die den funktionalen Wahrscheinlichkeitszusammenhang der betrachteten Konstrukte definiert, wie z.B. der von Personenfähigkeit und Aufgabenschwierigkeit (Thurstone 1925, Mosier 1941, Lazarsfeld 1973), siehe 2.2,
2) der Annahme der lokalen stochastischen Unabhängigkeit der Itembeantwortungen (u.a. Thurstone, 1926, Lazarsfeld 1966, Lord, Novick 1968, Lord 1980), siehe 2.4, und
3) der Schätzung der Parameter, wie z.B. der Schwierigkeit der verwendeten Testaufgaben (Items) und der Fähigkeit der Testpersonen, mit Hilfe des Maximum-Likelihood-Verfahrens (Fisher 1925, Fisher 1967, Kendall, Stuart, 1951), siehe 2.6.

© Springer Fachmedien Wiesbaden GmbH 2018
C. Fuhrmann, *Die trigonometrische Parametrisierung von Kompetenzen*, DOI 10.1007/978-3-658-19241-9_2

Das von Georg Rasch in den 60er Jahren des 20. Jahrhunderts entwickelte Item-Response-Modell geht von diesen Prämissen aus. Entscheidend für die Bedeutung des Rasch-Modells innerhalb der Testtheorie ist die Einbettung der Prämissen in das programmatische Konzept der spezifischen Objektivität, d.h. der Forderung der Unabhängigkeit des Vergleichs von Messungen vom verwendeten Messinstrument (Rasch 1960, Rasch 1966) – siehe 1.6 und 3.1 – und der der Unabhängigkeit der Eichung des Messinstruments von der verwendeten Eichstichprobe. Bisher erfüllt nur die logistische Item-Response-Funktion mit zwei Parametern, z.B. dem Aufgaben- und dem Personenparameter, diese Bedingung, siehe 3.1.2.

Während mit dem Begriff des Rasch-Modells die angesprochenen Merkmale eine auf stochastischen Grundsätzen beruhenden Item-Response-Funktion in das Konzept der spezifischen Objektivität einordnen, bezieht sich der Begriff der klassischen Testtheorie als Sammelbegriff auf alle diejenigen testtheoretischen Prinzipien, die nicht mit Antwortwahrscheinlichkeiten, sondern mit der Vorstellung eines „true scores“, also einer wahren Antwort als Limeskategorie und des aus ihr abgeleiteten statistischen Prozedere verbunden sind. Diese für die Methodologie des Rasch-Modells wie überhaupt für jede stochastisch bestimmte Item-Response-Theorie zentrale Differenz legt zugleich seine formale Struktur fest.

2.1 Idee der Item-Response-Theorie in Abgrenzung zur klassischen Testtheorie

Während sich die klassische Testtheorie mit dem Verhältnis von Testergebnis, Score, und „true score“ beschäftigt, nimmt die Item-Response-Theorie die Lösungen der einzelnen Items in den Blick. Sie geht davon aus, dass die Lösungen der Items lediglich durch *eine* latente Variable/Dimension/trait verursacht werden und definiert für jedes Item eine Item-Response-Funktion, die den funktionalen Zusammenhang zwischen der Wahrscheinlichkeit der Lösung eines Items und der latenten Variable bestimmt.

Die klassische Testtheorie geht von der Annahme aus, dass für jede Person und für jeden Test ein wahrer Score existiert, der die Fähigkeits- oder Einstellungsstruktur der Person widerspiegelt, der „true score“. Da jede Messung mit Fehlern behaftet ist, ist dieser true score bei einer einzelnen Messung grundsätzlich nicht ermittelbar. Erst bei unendlicher Wiederholung des Tests stellt sich im Mittel über alle Wiederholungen der true score ein, nämlich als Erwartungswert

der Messung[46]. Dies setzt voraus, dass der Messfehler zufällig verteilt und der Erwartungswert des Messfehlers Null ist (Fischer 1974: 27, Wieczerkowski, Schümann 1978: 44). An dieser Stelle werden zwei Probleme von Messungen deutlich:

1) Eine Messung ist grundsätzlich fehlerbehaftet und die Größe des Fehlers ist oft unbekannt (Harvill 1991, Traub, Roley 1991, Rost 2004: 36).
2) Erst in der unendlichen Wiederholung eines Tests/physikalischen Experiments ergibt sich der true score. (Harvill 1991)

Während die Messfehlerproblematik von Messungen für die Naturwissenschaften wie für die Erziehungs- und Sozialwissenschaften gilt, stellt gerade der zweite Punkt für die Erziehungs- und Sozialwissenschaften ein besonderes Problem dar. Zwar kann auch ein physikalisches Experiment trivialerweise nicht unendlich oft wiederholt werden, doch es kann sehr oft unter den immer wieder gleichen Bedingungen wiederholt werden.

In den Sozial- und Erziehungswissenschaften ändern sich mit jeder Testwiederholung die Bedingungen des Experiments: Sei es durch die Ermüdung der Testanden oder Auswirkungen, die die Konzentration stören oder Test-Lerneffekte.

Besonders problematisch sind systematische Fehler, wie z.B. durch Test-Lerneffekte. Wirken sich diese auf die Fähigkeit aus, die dem Lösen der Items zugrunde liegt, so verändert sich bei jeder Testwiederholung der true score der Untersuchungspersonen (Thussbas 2001: 23, Helmke 2004: 23, Minnameier 2005: 24, Pekrun 2002, Waldmann, Sydow 2006: 15, Embretson 1990: 423, Gruber/Mandl 1996: 589). Eine Wiederholung des Tests wäre in diesem Fall für die Bestimmung des true scores wirkungslos, da jede Testwiederholung etwas anderes, nämlich den durch den Lerneffekt veränderten „neuen" true score, messen würde.[47]

Die klassische Testtheorie postuliert folgende Zerlegung des true scores:

$$S = T + E \quad , \tag{5}$$

46 Definition des Erwartungswerts siehe z.B Bortz 1999, Tiede, Voss 2000, Wirtz, Nachtigall 2004, Dietz, Kalof 2010)

47 Auf die Problematik der Fehlergröße *einer* Messung, die sich implizit aus den beiden Punkten ergibt, wird noch genauer eingegangen. (siehe unter anderem 3.1)

mit *T* als dem true score, *S* dem Score/Testergebnis und *E* als dem (Mess-)Fehler (Error).

Die Zufälligkeit der angenommen Messfehlerentstehung verlangt, dass neben dem Erwartungswert des Messfehlers auch die Korrelation zwischen Messfehler und wahrem Wert Null ist.[48] Entsprechendes gilt für die Korrelation zwischen den Messfehlern unterschiedlicher wahrer Werte wie auch für die Messfehler von Tests, die das Gleiche messen. (Wieczerkowski, Schümann 1978: 45f, Fischer 1974: 31, Rost 2004: 36ff, Arnold, Nolda, Nuissl 2001: 110).

Da eine unendliche Wiederholung eines Tests nur definitorisch, aber nicht empirisch möglich ist, muss die Zuverlässigkeit einer durchgeführten Messung abgeschätzt werden (Reliabilität des Messinstruments/der Messung, siehe 1.6).

Für die klassische Testtheorie wird die Reliabilität eines Tests wie folgt definiert:

$$Rel = \frac{var(T)}{var(S)} = \frac{var(T)}{var(T) + var(E)} \leqslant 1 \quad , \tag{6}$$

mit *var(T)* als der Varianz der true scores und *var(S)* als der Varianz der Scores des Tests.[49]

Diese Definition gibt das Verhältnis der Varianz des wahren Werts im Verhältnis zur Varianz der gemessenen Werte an. Ist eine Messung völlig fehlerfrei, so sind beide Varianzen identisch, und es ergibt sich eine Reliabilität von Eins (Traub, Roley 1991).

Die Reliabilität kann aber auch als das Quadrat der Korrelation *r* zwischen dem wahren Wert *T* und dem gemessenen Score *S* verstanden werden:

$$Rel = \frac{var(T)}{var(S)} = \frac{cov(S,T)}{var(S)} = r^2(S,T) \quad , \tag{7}$$

mit $r^2(S,T)$ als dem Determinationskoeffizienten von *(S,T)* – Quadrat des Regressionskoeffizienten (Fischer 1974: 38). Auch hier wird die Kenntnis des wahren Werts zur Berechnung der Reliabilität benötigt.

Es ist aber auch möglich, die obige Korrelation, in der der unbekannte „true score“ verwendet wird, auf zwei Tests S_1 und S_2, die das Gleiche messen (Para-

48 Definition der Korrelation siehe z.B Bortz 1999, Wirtz, Nachtigall 2004, Rohwer, Pötter 2002.

49 Zur Definition der Varianz siehe z.B Bortz 1999, Tiede, Voss 2000, Wirtz, Nachtigall 2004, Rohwer, Pötter 2002.

lleltestmethode), zurückzuführen, da unter der Annahme, dass $var(E_1) = var(E_2)$, mit E_1 als dem Messfehler von S_1 und E_2 als dem Messfehler von S_2, gilt

$$Rel(S_1) = Rel(S_2) = r(S_1, S_2) \quad \text{(Rost 2004: 377).}^{50} \tag{8}$$

Aus der so berechneten Reliabilität lässt sich eine Schätzung der Messfehlervarianz angeben und damit ein Konfidenzintervall für den wahren Wert ableiten, denn wegen $\frac{var(T)}{var(S_1)} = r(S_1, S_2)$ und $\frac{var(E)}{var(S_1)} = 1 - r(S_1, S_2)$, da $var(S_1) = var(T) + var(E)$, ergibt sich:

$$var(E) = var(S_1)\sqrt{1 - r^2(S_1, S_2)} \quad \text{(Fischer 1974: 40).}^{51} \tag{9}$$

Auch ohne auf die methodischen, statistischen und inhaltlichen Probleme der Reliabilitätsmessung, wie z.B. das Reliabilitäts-Validitätsdilemma (Rost 2004: 392), dem Problem bei Veränderung der Heterogenität des untersuchten Merkmals innerhalb der Testgruppe (Fischer 1974: 38, Wieczerkowski, Schümann 1978: 47) oder der Annahme der Normalverteilung der Fehler (Wieczerkowski, Schümann 1978: 55) einzugehen, wird deutlich, dass sich die Berechnungen der klassischen Testtheorie immer am Score und an den Gruppenmittelwerten orientieren. Denn die Kennwerte, die zur Berechnung der Reliabilität notwendig sind, benötigen immer die Kenntnis des Mittelwerts der Untersuchungsgruppe.[52]

Die Item-Response-Theorie stellt hingegen die Wahrscheinlichkeit einer Itemlösung in den Mittelpunkt der Betrachtung. Für die Parameterberechnungen wird die Kenntnis eines Mittelwerts nicht benötigt.

2.2 Item-Response-Funktion

Durch den Wechsel der Perspektive vom Gesamtscore eines Test hin zur Analyse des itemabhängigen Verhaltens der Testpersonen definierten Tucker (1946: 4) und Lazarsfeld (1966: 81ff) die Konzepte der „item characteristic curve“ und der

50 Da $cov(S_1, S_2) = cov(T_1, T_2)$, denn der Erwartungswert der Fehler wird als Null angenommen.

51 Da $var(S) = var(T) + var(E)$ und die Korrelation und somit auch die Kovarianz zwischen T und E wie angenommen Null ist.

52 Zur Definition von Korrelation und Varianz siehe z.B Bortz 1999, Tiede, Voss 2000, Wirtz, Nachtigall 2004, Rohwer, Pötter 2002.

„traceline“. Beide beschreiben den funktionalen Zusammenhang zwischen der Ausprägung der Personenfähigkeit auf der zu untersuchenden Dimension und der Lösungswahrscheinlichkeit der betrachteten Aufgaben, also die schon genannte Item-Response-Funktion (bei Lord zunächst in Anlehnung an Tucker weiter Itemcharakteristik genannt [1953], ab 1968 wechselt die Bezeichnung dann zu Item-Response-Funktion).

Die Idee der Item-Response-Funktion geht auf Thurstone (1925) zurück. Die Lösung einer Aufgabe ist nach Thurstone durch einen funktionalen Zusammenhang der Ausprägung der Person auf dem trait (latente Eigenschaft) und der Schwierigkeit der Aufgabe gekennzeichnet.

Thurstone betrachtete dazu den von Binet entwickelten Test zur Messung von Intelligenz und analysierte die relativen Lösungshäufigkeiten einzelner Altersgruppen für einzelne, zufällig gewählte Testaufgaben des Binetschen Tests.

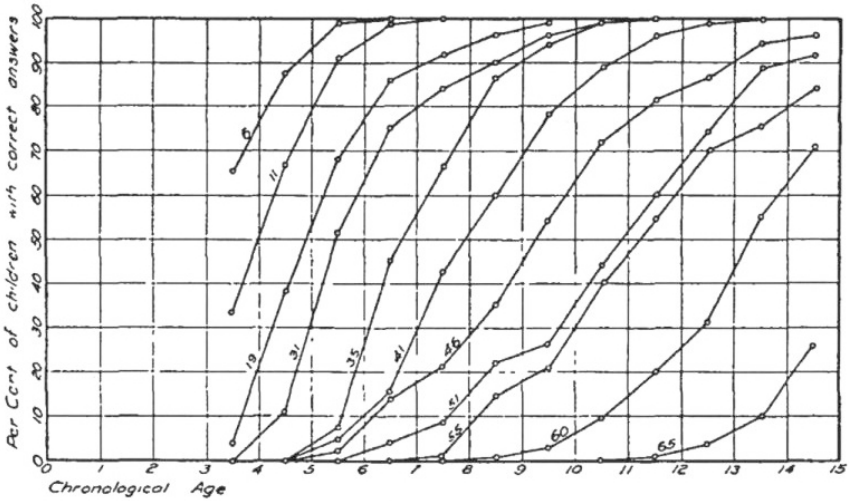


Abbildung 9: L.L. Thurstone, A Method of Scaling Psychological and Educational Tests, Journal of Educational Psychology. 1925: 444

Die Abszisse der Abbildung 9 repräsentiert das Alter der untersuchten Kinder, die Ordinate gibt die relative Häufigkeit der Lösung der untersuchten Aufgaben

an. Die Zahlen an den Verläufen der relativen Häufigkeiten geben an, welcher Testfrage der Verlauf zuzuordnen ist.

Man erkennt aus Abbildung 9, dass die relativen Lösungshäufigkeiten für alle untersuchten Aufgaben mit zunehmendem Lebensalter der untersuchten Kinder steigen. D.h. im Vergleich mit jüngeren Altersgruppen werden von älteren Kindern alle Aufgaben mit zunehmender Häufigkeit gelöst.

Aber es ist auch ersichtlich, dass die Aufgaben einer Rangordnung unterliegen. Zum Beispiel wird Aufgabe 6 schon von dreijährigen Kindern zu ca. 65% gelöst. Ab einem Alter von 6 Jahren lösen alle Kinder diese Aufgabe. Dagegen wird die Aufgabe 65 selbst von Kindern im Alter von 10 Jahren nicht gelöst. Auch in der Gruppe der 14-Jährigen schafft es nur etwas über ein Viertel der Kinder die Aufgabe zu lösen. Die anderen Aufgaben ordnen sich entsprechend.

Ausgehend von diesen Ergebnissen lässt sich der empirische Schwierigkeitsbegriff definieren:

Die Schwierigkeit oder Leichtigkeit von Aufgaben ergibt sich nicht durch die a priori Annahmen der Aufgabenentwickler, sondern durch die empirisch vorgefundenen Lösungsverhältnisse bei den Teilnehmern: Aufgaben gelten als leicht, wenn sie von vielen Testteilnehmern gelöst werden, sie gelten als schwierig, wenn nur wenige sie lösen (Sacher 1996: 172, Rohwer, Pötter 2002: 262f, Diekmann 2001: 238f, siehe 1.4). Im Beispiel von Thurstone ist damit Aufgabe 65 schwerer als Aufgabe 6.

Wird diese Schwierigkeitsdefinition akzeptiert, ergeben sich folgende Implikationen:

1) Leistungsfähige Testteilnehmer können schwere und leichte Aufgaben lösen. Sie lösen Aufgaben, die häufig gelöst wurden und solche, die nur von wenigen gelöst wurden. Weniger gute Teilnehmer lösen hingegen nur leichte Aufgaben, d.h. Aufgaben die von vielen gelöst wurden.
2) Werden in einem Test Aufgaben unterschiedlichen Schwierigkeitsgrades verwendet (leichte, mittlere und schwere Aufgaben), so kann die Leistungsfähigkeit der Untersuchungspersonen auch anhand der *Anzahl* der gelösten Aufgaben ermittelt werden. Denn gute Studierende werden eine höhere Anzahl von Aufgaben gelöst haben als weniger gute Studierende, da die Erstgenannten in der Lage sind leichte, mittlere und schwere Aufgaben zu lösen. Wenigr gute Studierende können nur leichte Aufgaben lösen. Übertragen auf die Darstellung der Abbildung 9 bedeutet dies, dass ältere Kinder mehr Aufgaben lösen als jüngere Kinder.

Die zweite Annahme ist jedoch keine deterministische Bedingung. Da es durchaus möglich ist, dass einige schwierige Aufgaben durch leistungsschwache Untersuchungspersonen erraten werden oder dass sie tatsächlich über das Wissen zur Lösung einiger schwerer Aufgaben verfügen. Vice versa können „gute" Untersuchungspersonen an leichten und schwierigen Aufgaben scheitern. Diese Unsicherheit kann durch die Wahrscheinlichkeit der Lösung einer Aufgabe beschrieben werden.

Daher stellen die in der Abbildung 9 dargestellten Verläufe nur den deskriptiven Zusammenhang der geforderten Beziehung von Lösungswahrscheinlichkeit einer Aufgabe und der ihr zu Grunde liegenden Dimension (trait) dar.[53]

Thurstone, Tucker, Lazarsfeld, Mosier[54] definieren den funktionalen Zusammenhang der Lösungswahrscheinlichkeit einer Aufgabe zunächst mittels der Differenz von Personenfähigkeit θ und Aufgabenschwierigkeit σ, $\theta - \sigma$, basierend auf dem von Thurstone 1926 definierten „Law of comparative Judgement". Diese Differenz $\theta - \sigma$ bestimmt die Wahrscheinlichkeit der Lösung einer Aufgabe durch eine Person

$$\theta_k - \sigma_l \quad , \tag{10}$$

mit θ_k als der Ausprägung der Fähigkeit von Person k auf der zu untersuchenden Fähigkeit und σ_l als der Schwierigkeit der Aufgabe l auf dem durch die Fähigkeit gebildeten „trait". Dieser Zusammenhang ist unmittelbar einsichtig denn es bedeutet nichts anderes, als dass die Lösungswahrscheinlichkeit einer Aufgabe zunimmt je größer die Differenz zwischen Personenfähigkeit und Aufgabenschwierigkeit ist.

Dabei geht die Testtheorie, und zwar sowohl in der klassischen Form wie auch in der Form der Item-Response-Theorie, davon aus, dass der Score, als die Summe der richtig gelösten Aufgaben/Items eines Tests, einen zentralen Indikator für die Ausprägung der Person auf der gemessenen (latenten) Eigenschaft darstellt (Fischer 1974: 180ff, Lord 1953: 517ff, 1980: 11, Lord, Novick 1968: 360).

53 siehe auch die Darstellung deskriptiver Item-Response-Funktionsverläufe in Harney, Fuhrmann 2010

54 Der Ansatz von Mosier (1941) unterscheidet sich dahingehend, dass Mosier ein „wahres" Antwortmuster für jede Person definiert und dann die Differenz dieses „wahren" Antwortmusters vom empirisch vorgefundenen Antwortmuster der Person analysiert: Sozusagen den Fehler der Abweichung der Antwortmuster.

Der Score ist

$$s_k = \sum_{l=1}^{N_I} x_{kl} \quad . \tag{11}$$

und entsprechend der Aufgabenscore t_l einer Aufgabe l

$$t_l = \sum_{k=1}^{N_P} x_{kl} \quad , \tag{12}$$

mit x_{kl} als der richtigen (als Eins kodiert) oder falschen (als Null kodiert) Lösung der Person k an der Aufgabe l, N_P als der Anzahl der Personen und N_I als der Anzahl der Aufgaben (Rost 2004: 91f).

2.3 Modelle der Item-Response-Theorie

Zur Modellierung der Wahrscheinlichkeit als Funktion der Differenz der Parameter wählte Tucker (1946: 4) die kumulative Dichtefunktion (Häufigkeitsverteilung) der Gauß-Verteilung (siehe unter anderem van der Linden, Hambleton 1997;5, weitere Formalisierung bei Lord 1953). Die Differenzannahme von Personenfähigkeit und Itemschwierigkeit in Verbindung mit der gaußschen Häufigkeitsverteilung wird in der Item-Response-Theorie als das Normal-Ogiven Modell bezeichnet:

$$p(x_{kl}=1|\theta_k, \sigma_l) = \int_{-\infty}^{a_l(\theta_k-\sigma_l)} \frac{1}{\sqrt{2\pi}} \exp(\exp(-t^2/2)\,dt) \quad . \tag{13}$$

Gl. (13) stellt die Wahrscheinlichkeit der richtigen Lösung von Aufgabe l der Person k dar, kodiert mit $x_{kl}=1$ für die richtige Lösung der Aufgabe l von Person k, mit θ_k als der Ausprägung der Person k auf der untersuchten latenten Fähigkeit, σ_l als der Schwierigkeit der Aufgabe k und a_l als der Steigung/Diskrimination des Verlaufs des Item-Response-Funktion. Das Integral selbst stellt dabei die kumulierte Gaußsche Dichtefunktion/Häufigkeitsverteilung dar.[55]

55 Definitionen der Gaußschen Verteilungsfunktion siehe z.B Bortz 1999, Tiede, Voss 2000, Wirtz, Nachtigall 2004, Rohwer, Pötter 2002, Dietz, Kalof 2010.

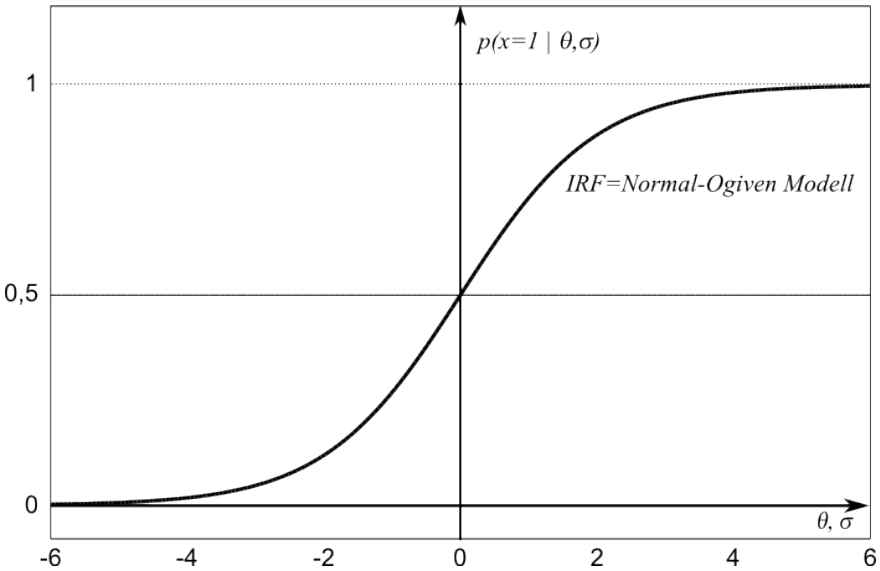


Abbildung 10: Visualisierung von Gl. (13), Darstellung der kumulativen Gaußschen Dichtefunktion/Häufigkeitsverteilung als Item-Response-Funktion in Abhängigkeit der Differenz von θ und σ

Abbildung 10 verdeutlicht in anschaulicher Weise das Prinzip der Item-Response-Theorie: Je größer die Differenz aus Personenfähigkeit und Aufgabenschwierigkeit ist, desto größer ist die Wahrscheinlichkeit die betreffende Aufgabe zu lösen.

Dabei stellt die Wahl der gaußschen Häufigkeitsverteilung zur Modellierung der Differenz von Personenfähigkeit und Aufgabenschwierigkeit keine statistische Notwendigkeit dar, sondern ist nur eine von vielen möglichen sigmoiden Funktionen (Baker, Seock-Ho 2004: 6, Spada 1976: 91), die einen Wertebereich von Null und Eins aufweisen. Das Letztere ist eine Anforderung, die an eine Funktion gestellt werden muss, die Wahrscheinlichkeiten als Ergebnis darstellen soll.[56] Die Begründung der Wahl der normal Ogive zur Modellierung der Diffe-

56 Werte für Wahrscheinlichkeiten liegen per Definition zwischen Null und Eins, wobei Null für ein Ereignis steht, dass mit Sicherheit nicht eintritt und Eins für ein Ereignis, das auf jeden Fall eintreten wird.

renz von Personenfähigkeit und Aufgabenschwierigkeit zur Berechnung der entsprechenden Lösungswahrscheinlichkeit wird mit dem besonders plausiblen Verlauf der gaußschen Häufigkeitsverteilung begründet (Rost 2004: 115, Baker, Seock-Ho 2004: 6, Lord, Novick 1968: 367, Lord 1953: 524, Birnbaum in Lord, Novick 1968: 400, van der Linden, Hambleton 1997: 5).

In der Entwicklung von Item-Response-Modellen wurde zunächst das Normal-Ogiven Modell mit den beiden Parametern θ und σ durch zusätzliche Parameter erweitert, wie z.B. Parametern für unterschiedlich diskriminierende Items, $a_l \cdot (\theta_k - \sigma_l)$ (2-pl-Modell[57]) oder Modelle, die Rateparameter für die Möglichkeit des Erratens der richtigen Lösung in ihr Modell integrieren (3-pl-Modell siehe Gl. (14) und Abbildung 11):

$$p(x_{kl}=1|\theta_k, \sigma_l, c_l, a_l) = c_l + (1-c_l) \int_{-\infty}^{a_l(\theta_k-\sigma_l)} \frac{1}{\sqrt{2\pi}} \exp((-t^2/2)\,dt) \quad , \quad (14)$$

mit c_l als dem Rateparameter und a_l der Steigung/Diskrimination der Aufgabe l (Lord 1970: 45, 1975: 205).

57 Die numerischen Bezeichnungen der Modelle ergeben sich aus der Anzahl der für die Aufgaben zu schätzenden Parameter. Für das 2-pl-Modell sind dies die Aufgabenschwierigkeit und die Diskrimination je Item.

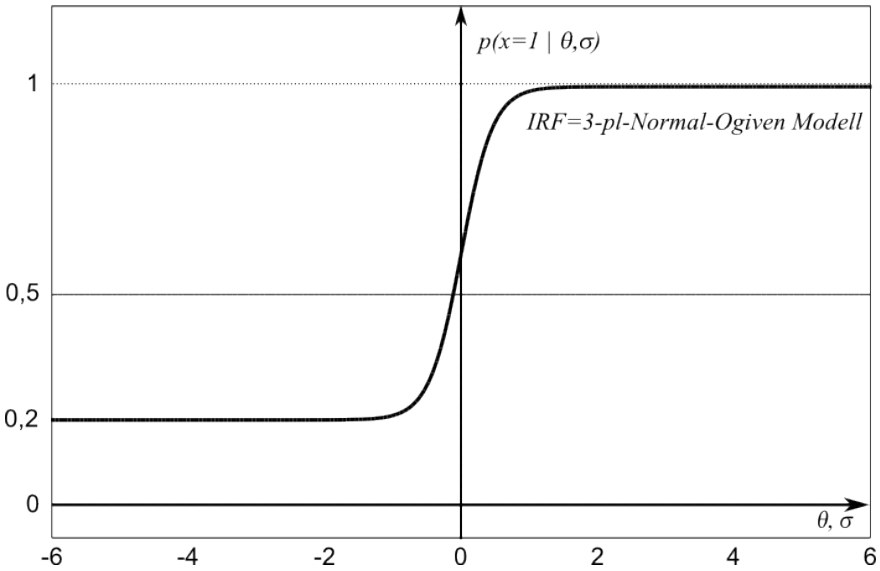


Abbildung 11: Darstellung des Verlaufs einer Item-Response-Funktion des 3-pl-Modells mit dem Rateparameter $c = 0{,}2$, dem Diskriminationskoeffizienten $a = 1$ und der Itemschwierigkeit $\sigma = 0$

Abbildung 12 zeigt den Verlauf zweier Item-Response-Funktionen des normalogiven Modells mit unterschiedlich ausgeprägten Koeffizienten a_l, c_l und σ_l.[58]

58 Abbildung 11 und Gleichung (14) verdeutlichen, dass im Falle von $c = 0$ aus dem 3-pl-Modell das 2pl-Modell entsteht.

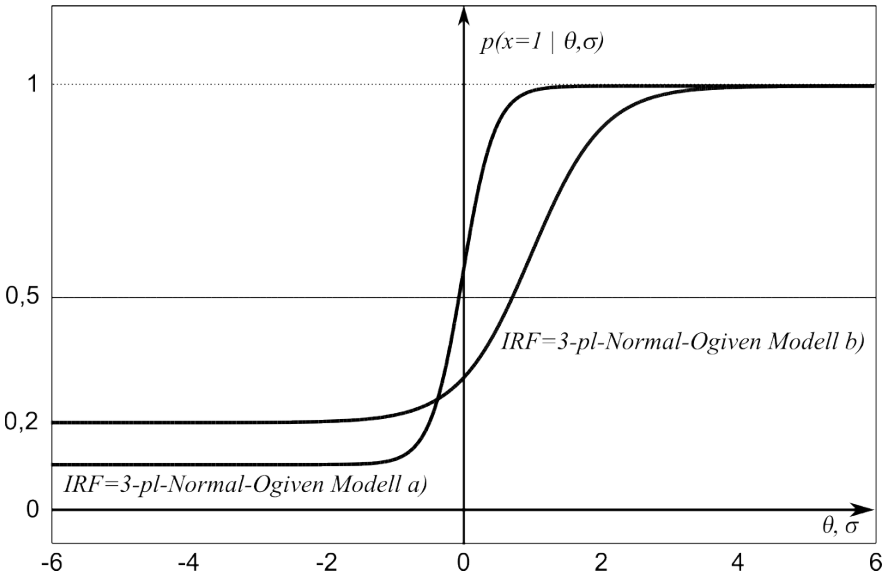


Abbildung 12: Exemplarische Darstellung zweier Normal-Ogiven 3-pl-Item-Response-Funktionen mit unterschiedlichen Rate- und Diskriminationsparametern: Item-Response-Funktion a) stellt einen Funktionsverlauf mit *c = 0,2, a = 4,3* und Schwierigkeitsparameter Null, Item-Response-Funktion b) einen mit Funktionsverlauf mit c *= 0,1, b = 1,9* und Aufgabenschwierigkeit Eins.

Per Definition wird die Itemschwierigkeit (location) an die Stelle der Wendestelle[59] der Item-Response-Funktion gelegt. Im Falle der gaußschen Verteilungsfunktion ist dies der Punkt bei dem die Funktion den Wert 0,5 annimmt. Dieser Punkt repräsentiert die Gleichwahrscheinlichkeit von Lösung und Nicht-Lösung der Aufgabe.[60]

59 An der Wendestelle ist die zweite Ableitung einer Funktion gleich Null, siehe z.B. Forster 1985. An dieser Stelle ändert sich die Krümmung einer Funktion.

60 Auch die betrachtete logistische Item-Response-Funktion hat bei der Wahrscheinlichkeit 0,5 ihre Wendestelle (siehe die sich direkt im Anschluss befindliche Abbildung 13), wie auch die in dieser Arbeit entwickelte trigonometrische Item-Response-Funktion. In den PISA Auswertungen wurde die Itemlokation an die Stelle mit der Lösungswahrscheinlichkeit 0,62 gelegt (Adams, Wu 2002: 207). Die Idee ist, dass

Georg Rasch (1960) und Birnbaum (in Lord, Novick 1968) verwendeten die logistische Funktion als Item-Response-Funktion und definierten so die logistischen Modelle der Item-Response-Theorie, z.B. das logistische Modell mit Rate- und Diskriminationsparameter:

$$p(x_{klk}=1|\theta_k, \sigma_l, c_l, a_l) = c_l + \frac{1-c_l}{1+\exp(-a_k(\theta_k-\sigma_l))} \quad [61] \qquad (15)$$

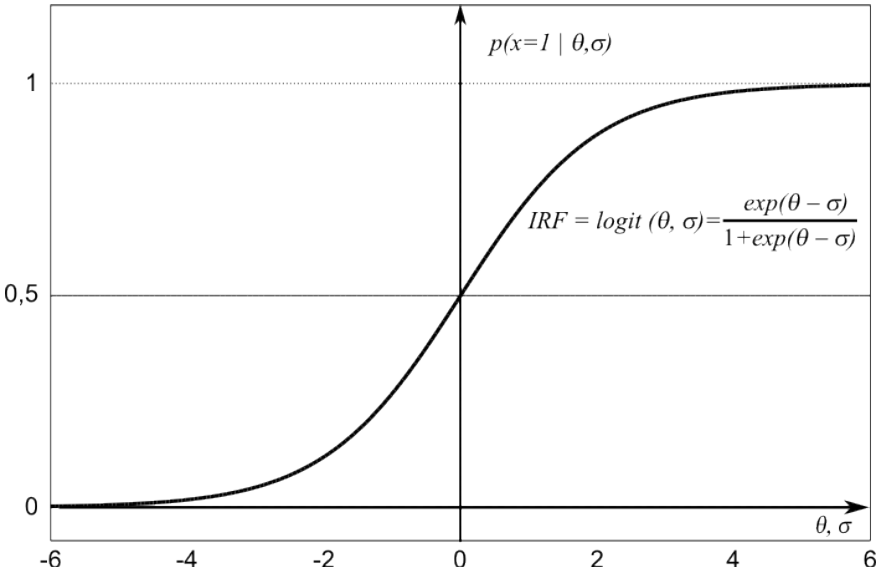


Abbildung 13: Logistische Item-Response-Funktion mit $c = 0$, $a = 1$ und $\sigma = 0$

Als Begründung für die Verwendung der logistischen Funktion an Stelle des Normal-Ogiven Modells werden deren Ähnlichkeit im Verlauf angeführt. Durch

durch diese Verschiebung die teilnehmenden Schüler und Schülerinnen gesicherter den postulierten Kompetenzstufen zugeordnet werden können (Adams, Wu 2002: 197).

61 Das negative Vorzeichen innerhalb der Exponentialfunktion verschwindet, wenn die Modellgleichung, wie üblich, als Term angegeben wird der sowohl das Ereignis $x_{kl} = 0$ wie auch das Ereignis $x_{kl} = 1$ darstellen kann (Rost 2004: 133ff)

die Einführung eines Faktors *D=1,701* (Lord 1975: 205, Birnbaum 1968 in Lord, Novick 1968: 400, Samejima 2000: 319, Swaminathan 1983: 25) werden die beiden Verläufe der Funktionen noch weiter aneinander angepasst:

$$p(x_{kl}=1|\theta_k, \sigma_l, c_l, a_l)=c_l+\frac{1-c_l}{1+\exp(-D\cdot a_l(\theta_k-\sigma_l))} \quad . \tag{16}$$

Die Verwendung der logistischen Funktion als Item-Response-Funktion vereinfacht die mathematische Berechnung der Parameter des Modells (Bock, Moustaki 2007: 476, Bejar 1983: 12). Diese Vereinfachung war entscheidend für die Anwendung und Verbreitung der Item-Response-Theorie, denn die anzuwendenden numerischen Verfahren zur Bestimmung der Parameter waren zwar bekannt, doch standen für ihre algorithmische Implementierung auf Computern zur Entwicklungszeit des Rasch-Modells keine ausreichenden technischen Rechenkapazitäten bereit (Samejima 2000: 320) bzw. waren nur sehr teuer zu beschaffen. Die Vereinfachung der Parameterberechnung schuf daher die Grundlage dafür, ein Modell überhaupt berechnen zu können.

Das Rasch-Modell stellt für Birnbaum nur einen Spezialfall seines allgemeinen Modells dar, nämlich das 1-pl-Modell, Abb. 13, da es keinen Rateparameter c_k kennt und die Aufgaben alle den gleichen Diskriminationskoeffizienten haben, üblicherweise $a = 1$ für alle Items. Somit ist für die Aufgaben jeweils nur der Schwierigkeitsparameter zu schätzen (van der Linden/Hambleton 1997: 8).

Dem von Birnbaum definierten allgemeineren Modell fehlt jedoch die Eigenschaft spezifisch objektiv zu sein. Spezifische Objektivität setzt nämlich voraus, dass die Items gleiche Steigung aufweisen (Fischer 1974: 174, siehe Abbildung 4).

Auch das Rasch-Modell selbst wurde variiert, z.B. durch das linear logistische Test Modell (LLTM) nach Fischer (1974: 340ff, 1995, 1995a):

$$p(x_{kl}=1|\theta_k, \sigma_l=\sum_{i=1}^{M} a_{il}\beta_i)=\frac{\exp(\theta_k-\sum_{i=1}^{M} a_{il}\beta_i)}{1+\exp(\theta_k-\sum_{i=1}^{M} a_{il}\beta_i)} \quad . \tag{17}$$

Gl. (17) stellt die Modellgleichung des LLTM für eine Person k bzgl. einer Aufgabe l dar. Dabei wird die Aufgabenschwierigkeit σ_l in die zu ihrer Lösung notwendigen kognitiven Teilfähigkeiten β_i zerlegt und zusätzlich mit einem Gewich-

tungsfaktoren je Teilfähigkeit a_{il} versehen. Für die Teilfähigkeiten wird eine linear additive Beziehung angenommen, ähnlich wie in der multiplen Regressionsrechnung.

Auf die die Parameter noch weiter differenzierenden Modellentwicklungen von Embretson und Dibello (Embreston 1984, 1990, Dibello, Roussos, Stout 2007) wird hier nicht weiter eingegangen. Gemein ist aber all diesen Modellen, dass sie versuchen unterschiedliche Item- und Kognitionsstrukturen zu berücksichtigen oder mehrdimensionale Fähigkeitskonstrukte abzubilden, und dass sie nicht spezifisch objektiv sind.

Alle hier genannten Modelle der Item-Response-Theorie setzen allerdings die erwähnte lokale stochastische Unabhängigkeit voraus, die in Abschnitt 2.4 ausführlich dargestellt wird, und die Abhängigkeit der Parameter von den Scores.[62]

2.3.1 Kritik der dargestellten Modelle der Item-Response-Theorie

Die Verwendung des Scores als alleinige Informationsquelle zur Bestimmung der Fähigkeitsausprägung von Personen und Items wird in der Literatur von folgenden Autoren kritisiert:

Glaser (1985: 55ff) und Snow und Lohman (1989: 320ff) stützen ihre Kritik auf Beobachtungen aus der Experten- und Novizenforschung: Für sie enthalten auch die nicht gelösten Aufgaben essentielle Informationen für die Fähigkeitsbestimmung von Personen über dem Konstrukt. Tatsuoka setzte sich 1984 mit der Analyse der Antwortmuster der Personen auseinander und entwickelt daraus später das Rule Space Model (1990), siehe 4.5.

Auch Dibello, Roussos, Stout (2007: 981ff) und Wilson (2005: 28) weisen darauf hin, dass für die kognitive Diagnose selbst bei Vorliegen einer unidimensionalen Fähigkeitsstruktur wichtige Informationen in den erhobenen Daten enthalten sein können, die jenseits einer reinen Scorebetrachtung liegen. Daher fordern sie, dass neue Modelle der Item-Response-Theorie entwickelt werden müssen, um diese Informationen auswerten zu können. Leighton und Gierl (2002,

62 Fischer (1974: 228) gibt an, dass sein Modell nur vom Vektor der Randsummen der Daten abhängt und Embretson setzt die Suffizienz der Scores voraus, aufgesplittet nach der Summe richtiger oder falscher Beantwortungen (Embretson 1984: 181ff, siehe Darstellung der Suffizienz in Abschnitt 3.1.3).

2007) weisen darauf hin, dass erst eine Analyse der Antwortmuster der Personen Diagnostik überhaupt ermöglicht.

Die dazu bisher vorliegenden Ansätze zur Ausschöpfung weiterer Informationen aus vorliegenden Daten, wie z.B. die Ansätze von Samejima (2000, 2008), Reise (2000) oder Molenaar und Hoitjink (1990), lösen die Forderung nach entsprechenden Modellen aber lediglich partiell ein.

Reise und Molenaar und Hoitjink entwerfen „nur" geänderte Modellanpassungstest für das Rasch-Modell, bei denen besonderer Augenmerk auf Gruppen von Antwortmustern gelegt wird, die mit dem Rasch-Modell nicht verträglich sind (siehe 5.2.1.3). Samejima entwickelt ein Modell auf der Grundlage der logistischen Item-Response-Theorie, in dem die Item-Response-Funktionen nicht punktsymmetrisch sind. D.h., die Wendestelle der Funktionen liegt nicht wie bei der logistischen Funktion bei einem Funktionswert von 0,5 (siehe eben), sondern darüber oder darunter. Das so entstehende Modell ist in der Lage, Antwortmuster zu identifizieren. Es wertet Person mit hohem Score ab, wenn diese Personen leichte Aufgaben nicht gelöst haben, und kreditiert Personen, die einen geringen Score aufweisen, aber dabei schwierige Aufgaben lösen konnten. Die Berechnung der Parameter bleibt jedoch weiterhin ausschließlich durch den Score determiniert und diese Modellfamilie erfüllt nicht den besonderen messtheoretischen Ansatz des Rasch-Modells (siehe 3.1 und 1.2 und 1.6).[63]

Nach dieser ersten Einschätzung von Modellen der Item-Response-Theorie werden nun die allen Item-Response-Modellen zugrunde liegenden Annahmen erörtert.

2.4 Lokale stochastische Unabhängigkeit

Das Konzept der lokalen stochastischen Unabhängigkeit, das von Lazarsfeld (1973: 369)[64] zunächst "definition of a pure test" genannt und später als „local

63 Nur das Rasch-Modell ist, nach der von Rasch und später von Fischer verallgemeinerten Definition der spezifischen Objektivität und bei Verwendung einer monotonen Item-Response-Funktion, spezifisch objektiv, siehe Fischer 1974: 407, 1989: 87 und 1994: 27.

64 Die Entwicklungen und Ideen von Lazarsfeld (1973) zeigen auch einen entscheidenden Unterschied zu dem von Lord ab 1953 entwickelten Modell der Item-Response-Theorie. Während es Lazarsfeld um die Analyse der Lösungsmuster (Antwortmuster) der untersuchten Personen ging, also die Analyse welche Person welche Aufgabe löst oder nicht löst, interessierte sich Lord in Anlehnung an die klassische Testtheorie nur noch für den Score der Personen, also die Anzahl der gelösten Aufgaben unabhängig

independence“ oder „conditional independence“ definiert wurde, ist einer der zentralen Begriffe der Item-Response-Theorie.

Anschaulich, aber verkürzt, besagt die Definition der lokalen stochastischen Unabhängigkeit, dass Items nicht über ihre Lösungen miteinander verkettet sein dürfen. Das bedeutet, dass für die Lösung eines Items B die richtige Lösung des Items A nicht Voraussetzung sein darf.[65]

Es kann jedoch davon ausgegangen werden, dass Items, die ein und die selbe latente Fähigkeit/Dimension beschreiben, über alle Personen hinweg miteinander korreliert sind, denn die Lösungen der Items sind abhängig von der Ausprägung der Personen auf der latenten Fähigkeit.

Werden die Itemlösungen hingegen nicht an allen Personen, sondern nur an Personen einer einzigen Fähigkeitsausprägung der latenten Eigenschaft betrachtet, dann sind die Items unkorreliert. D.h. die lokale Einschränkung (gegeben den Score) nimmt die Korrelation über die Funktion heraus, da diese nun nicht mehr variiert. Denn in diesem Fall ist das, was die Variation der Lösungen der Items beeinflusst und als Korrelation berechenbar ist, in der Personengruppe konstant, da alle die gleiche Fähigkeitsausprägung aufweisen. Wäre das Antwortverhalten in den Fähigkeitsgruppen deterministisch gegeben, so würde sich für jedes Gruppenmitglied das gleiche Lösungsmuster der gelösten/nicht gelösten Items ergeben. Die von diesem Lösungsmuster abweichenden empirischen Antwortmuster der Personen in den Fähigkeitsgruppen sind dann als stochastische Schwankungen aufzufassen. Die jeweilige Fähigkeitsgruppe, als Bedingung der Gültigkeit der stochastischen Unabhängigkeit, weist der stochastischen Unabhängigkeit den Ort, zu in dem sie gültig ist. Aufgrund dieser Bedingtheit der lokalen stochastischen Unabhängigkeit wird diese auch als „conditional independence“ bezeichnet (Bock, Moustaki 2007: 470).

Eine erste Definition der lokalen stochastischen Unabhängigkeit findet man in einem Aufsatz bei Lazarsfeld in „The logical and mathematical foundation of latent structure analysis“ (1973: 367ff, Neuauflage der Erstausgabe von 1950) und in einem Aufsatz (1966: 80). In beiden Publikationen beschäftigt sich Lazarsfeld mit Klassen von Personen, denen jeweils *klassenspezifisch* unterschiedliche „Fähigkeiten“ zukommen.

Die formale Definition der lokalen stochastischen Unabhängigkeit kann mittels zweier äquivalenter Ausdrücke erfolgen:

---

davon welche Aufgaben gelöst wurden.

65 Verkürzt deshalb, da die grundlegende Idee der lokalen stochastischen Unabhängigkeit an dieser Formulierung nicht sichtbar wird. Zudem trifft diese Formulierung auf alle voneinander unabhängigen stochastischen Prozesse zu.

$$p(x_i=1|\theta)=p(x_i=1|\theta, x_j, x_k, \ldots) \quad , i\neq j, k, \ldots \tag{18}$$

oder beispielhaft für drei Items:

$$p(x_i=1, x_j=1, x_k=1|\theta)=p(x_i=1|\theta)\, p(x_j=1|\theta)\, p(x_k=1|\theta) \tag{19}$$

für drei Items *i, j, k,* mit der Kodierung $x_i=1$, $x_j=1$, $x_k=1$ für Item *i, j, k* richtig gelöst, θ als einer beliebigen aber festen Personenfähigkeit und *p* als der Wahrscheinlichkeit des jeweiligen Ereignisses, z.B. $p(x_i=1|\theta)$ als der Wahrscheinlichkeit Aufgabe *i* bei gegebener Personenfähigkeit θ zu lösen: „x_i unter der Bedingung θ" (Lord 1980: 19).

Die Äquivalenz beider Definitionen ergibt sich aus den Rechenregeln für Wahrscheinlichkeiten und dem Satz von Bayes über bedingte Wahrscheinlichkeiten. Der Satz von Bayes besagt, dass

$$p(x_i=1, x_j=1|\theta) = p(x_i=1|\theta)\, p(x_j=1|\theta, x_i=1) \quad . \tag{20}$$

Zusammen mit der Forderung aus Gl. (18) ergibt sich dann Gl. (19) (Kendall, Stuart 1947: 181, Tiede, Voss 2000: 36, Rohwer, Pötter 2002: 81, Rost 2004: 160, Robert, Chopin, Rousseau 2009: 7).

Lord beschrieb dabei die Wahrscheinlichkeit der rechten Seite von Gl. (18) als die Wahrscheinlichkeit der richtigen Lösung des Items *i* bei einer Person mit der Fähigkeit θ und den gegebenen Lösungen dieser Person bezüglich aller anderen Items (Lord 1989: 19).

Lazarsfeld bezieht sich bei seiner Definition der lokalen stochastischen Unabhängigkeit auf Gl. (19). Durch Gl. (19) kann der erwähnte Eindruck entstehen, dass lokale stochastische Unabhängigkeit alleine in der Forderung aufgeht, dass die Lösungen von Items nicht verkettet sein dürfen. D.h. dass die Voraussetzung zur Lösung eines Items nicht die Kenntnis der richtigen Antwort eines vorher bearbeiteten Items sein darf. Vielmehr besagt die lokale stochastische Unabhängigkeit, dass die Lösung der Items unter der Bedingung einer konstanten Personenfähigkeit selbst ein stochastischer Prozess ist, d.h. dass die Lösung der Items alleine vom situativen, zufälligen Kontext bestimmt ist. Daher ist die „Nicht-Verkettung" der Aufgabenlösungen eine Folgerung, die sich aus dieser Forderung ableitet.

Georg Rasch geht in einem Aufsatz aus dem Jahr 1961 auf den Begriff der „conditional independence" ein (Rasch 1961: 325). Hier, wie auch in zwei Auf-

sätzen aus dem Jahr 1966, greift er das Problem der Korreliertheit der Lösungen der Items über der Personenfähigkeit auf (Rasch 1966: 55, 1966a: 92).

Seine Modellentwicklung geht davon aus, dass die Items nicht nur lokal stochastisch unabhängig sind, sondern unbedingt. Rasch zeigt, dass unter der Voraussetzung der stochastischen Unabhängigkeit und der Annahme, dass Personen gleichen Scores die gleiche Ausprägung auf der Fähigkeitsdimension haben, die Definition der bedingten stochastischen Unabhängigkeit von Lord und Lazarsfeld folgt, siehe auch Gl. (19) (Rasch 1966a: 93).

Zum Beleg, dass aus der Forderung der stochastischen Unabhängigkeit im Falle des Rasch-Modells die lokale stochastische Unabhängigkeit folgt, verwendet Rasch ein Beispiel für eine Person und zwei Items.

Unter der Annahme der stochastischen Unabhängigkeit gilt für die Wahrscheinlichkeit jeweils:

$$\begin{cases} p(x_i{=}1, x_j{=}1) = p(x_i{=}1)\, p(x_j{=}1) \\ p(x_i{=}1, x_j{=}0) = p(x_i{=}1)\, p(x_j{=}0) \\ p(x_i{=}0, x_j{=}1) = p(x_i{=}0)\, p(x_j{=}1) \\ p(x_i{=}0, x_j{=}0) = p(x_i{=}0)\, p(x_j{=}0) \end{cases} . \tag{21}$$

Betrachten wir nur die ersten beiden Zeilen von Gl. (21), d.h. die betrachtete Person hat entweder beide Items gelöst oder Item *i* gelöst aber Item *j* nicht, und dividieren die erste Wahrscheinlichkeit durch $p(x_j{=}1)$ und die zweite durch $p(x_j{=}0)$ so ergibt sich:

$$p(x_i{=}1) = \frac{p(x_i{=}1, x_j{=}1)}{p(x_j{=}1)} = \frac{p(x_i{=}1, x_j{=}0)}{p(x_j{=}0)} \quad . \tag{22}$$

Der linke Quotienten von Gl. (22) entspricht der Definition der bedingten Wahrscheinlichkeiten nach Bayes, siehe Gl. (20), :

$$p(x_i{=}1 | x_j{=}1) = \frac{p(x_i{=}1, x_j{=}1)}{p(x_j{=}1)} \quad . \tag{23}$$

Da die gleiche Bedingung für den rechten Quotienten für $x_j = 0$ gilt kann Gl. (20) geschrieben werden als

$$p(x_i{=}1 | x_j{=}1) = p(x_i{=}1 | x_j{=}0) = p(x_i{=}1) \tag{24}$$

Folglich ist die Lösungswahrscheinlichkeit von $p(x_i=1)$ für die betrachtete Person unabhängig von der Lösungswahrscheinlichkeit $p(x_j)$ und die Aufgaben sind somit stochastisch unabhängig, was für alle verwendeten Aufgaben gilt. D.h. aus der lokalen stochastischen Unabhängigkeit folgt die unbedingte stochastische Unabhängigkeit (Rasch 1966a: 93).

Die Korreliertheit der Items ergibt sich, wenn die Items nicht für eine Personenfähigkeit, sondern über das gesamte Spektrum der Personenfähigkeiten betrachtet werden.

Die Forderung der stochastischen Unabhängigkeit kann somit in die Definition der lokalen stochastischen Unabhängigkeit überführt werden und stellt keinen Widerspruch zur Korreliertheit der Items dar (Rasch 1961: 325).

Raschs Ausführungen zum Zusammenhang von stochastischer Unabhängigkeit und lokaler stochastischen Unabhängigkeit begründen sich durch seine Sichtweise auf den Prozess der Itembeantwortung: Für Rasch ist die Antwortreaktion auf ein Item ein grundsätzlich stochastischer Prozess, da das Lösungsverhalten von Personen nicht deterministisch vorhergesagt werden kann und zwar unabhängig davon, ob Personenfähigkeit und Aufgabenschwierigkeit miteinander korreliert sind (Rasch 1960: 10ff).[66]

2.5 Unidimensionalität

Notwendige Bedingung für die Existenz der lokalen stochastischen Unabhängigkeit ist, dass die Items die Forderung der Unidimensionalität erfüllen. Dies bedeutet, dass die Lösung der in einem Test verwendeten Items nur durch eine einzige latente Variable (z.B. Fähigkeit der Person) verursacht wird. Das Konzept der bedingten stochastischen Lösungswahrscheinlichkeit ist ohne die Unidimensionalität nicht denkbar. Erst durch die Unidimensionalität der Aufgaben werden die Antwortmuster aller Personen einer Fähigkeitsausprägung alleine durch unsystematische Einflüsse variiert. Andernfalls würden weitere latente Faktoren die Antwortmuster beeinflussen können. Erst die Unidimensionalität ermöglicht es die latente Eigenschaft durch die Einschränkung der Personen auf

66 Rasch vertritt damit die von den meisten Vertretern der Item-Response-Theorie gemachte Auffassung, dass sich der stochastische Prozess der Item-Response-Theorie durch das individuelle Verhalten der Personen manifestiert (Klauer 1995: 97). Eine andere Position ist, dass sich der stochastische Prozess durch die Ziehung der Zufallsstichprobe von Personen aus einem Grundgesamtheit bzw. von Aufgaben aus einem Itemuniversum einstellt.

eine Fähigkeitsausprägung „herauszupartialisieren“[67]. Zudem ist es erst dann sinnvoll Personenfähigkeit und Aufgabenschwierigkeit in eine Differenzbeziehung zu bringen, da nun erst die sich ergebende Aufgabenschwierigkeit und die sich einstellende Personenfähigkeit durch die gleiche Dimension hervorgerufen werden.

Es ist jedoch zu beachten, dass die Forderung der Unidimensionalität der Items zwar eine notwendige Bedingung des Modells darstellt, diese aber unabhängig von der Forderung der lokalen stochastischen Unabhängigkeit ist, wie Gl. (21) bis Gl. (25) zeigen. Denn die bedingte stochastische Unabhängigkeit kann alleine aus der stochastischen Unabhängigkeit der Lösungswahrscheinlichkeit der Items bzgl. einer Person abgeleitet werden (Stout 1987: 589).

Die Forderung der Unidimensionalität ist eine sehr restriktive Forderung und erschwert die Testkonstruktion. Die Konstruktion reiner unidimensionaler Items ist zeit- und kostenintensiv. Die Behauptung, dass Items eines Tests unidimensional sind, wird immer kritisierbar sein. Doch auch, wenn die grundsätzliche Kritik an der Behauptung, dass Items eines Tests unidimensional sind, berechtigt ist, kann davon ausgegangen werden, dass die Beantwortung/Lösung unidimensional konstruierter und getesteter Items zumindest durch die angenommene zu Grunde liegende Dimension dominiert wird (Sijtsma, Verwej 1999: 58).

2.6 Maximum-Likelihood: Parameterschätzung in der Item-Response Theorie

Im Folgenden wird mit der Maximum-Likelihoodlihood Methode ein Verfahren beschrieben, das aus den gegebenen Daten einer Stichprobe Parameter bestimmt, so dass die Stichprobe mit höchster Wahrscheinlichkeit aus einer Population stammt, in der die Daten entsprechend der berechneten Parametern verteilt sind.

Das bekannteste Beispiel eines aus einer Stichprobe geschätzten Parameters für die Population ist der Mittelwert einer Stichprobenvariablen. Der Mittelwert der Stichprobe schätzt den Mittelwert einer Population erwartungstreu/unbiased[68], wenn die Merkmalsausprägungen in der Population als gaußverteilt gelten.

67 vgl. Bortz 1999: 350.

68 Erwartungstreu bedeutet im Fall des Populationsmittelwerts μ, dass sich dieser Mittelwert als Erwartungswert der Stichprobenmittelwerte ergibt (Bortz 1999: 95). Kendall und Stuart geben zum Begriff der Erwartunsgtreue („unbiased“) die folgende Definition: Die Stichprobenverteilung eines Parameters ist unbiased, wenn für alle Stichprobengrößen und „wahren Werte” des Parameters der Erwartungswert des Parameters

Der Mittelwert der Stichprobe ist daher ein Schätzer für den Mittelwert der Population (Bortz 1999: 98).

Die Idee der Maximum-Likelihood-Methode ist nun die folgende: Gegeben seien die Daten einer Stichprobe. Welchen Wert müssen die Parameter der für die Population angenommen Verteilung annehmen, so dass die Daten der Stichprobe mit höchster Wahrscheinlichkeit/Likelihood aus der Population stammen (vgl. Mung 2003).

2.6.1 Ein einfaches Beispiel zur Illustration der Maximum-Likelihood-Methode

Anhand des Beispiels des Münzwurfs wird nachfolgend das Verfahren der Bestimmung von Verteilungsparametern mit Hilfe der Maximum-Likelihood-Methode illustriert:

Annahme:

Nach zehnmaliger Wiederholung eines Münzwurfs erhält man als Ergebnis siebenmal Kopf und dreimal Zahl. Die durch die Maximum-Likelihood-Schätzung zu beantwortende Frage ist:

Welche Wurf-Verteilung (Wahrscheinlichkeitsverteilung) hat die Münze aufgrund dieser Datenlage/Stichprobe am wahrscheinlichsten? Ist die Münze fair?

Die Wahrscheinlichkeit q eines einzelnen Münzwurfs lässt sich mit Hilfe der Binomialverteilung wie folgt darstellen (Rost 2004: 111)

$$q(x_k|\xi)=\xi^{x_k}\cdot(1-\xi)^{(1-x_k)} \quad , \tag{25}$$

wobei x_k das Ergebnis des Münzwurfs darstellt, d.h. Kopf oder Zahl kodiert als 1 oder 0, und ξ als dem einzigen Wurf-Verteilungsparameter der Münze, d.h. ξ gibt die Wahrscheinlichkeit an mit der das Ergebnis des Münzwurfs Kopf eintritt.[69]

Für die 10 Würfe ergibt sich eine Gesamtwahrscheinlichkeit der Stichprobe von

$$p(x|\xi)=\binom{10}{7}\prod_{k=1}^{10}\xi^{x_k}\cdot(1-\xi)^{(1-x_k)} \quad , \tag{26}$$

gleich dem wahren Wert ist (Kendall, Stuart 1951: 4, Dietz, Kalof 2009: 280).

69 Im Falle einer fairen Münze wäre der Wert von $\xi=0,5$, d.h. die Münze würde bei einer unendlichen Anzahl von Würfen ebenso oft Kopf wie Zahl zeigen.

denn die Ergebnisse der einzelnen Würfe sind voneinander unabhängig.[70] Somit ergibt sich die Gesamtwahrscheinlichkeit aller zehn Würfe als Multiplikation der Wahrscheinlichkeiten der Einzelwürfe (Kendall, Stuart: 1947: 180). Da das Ergebnis der zehn Münzwürfe nicht von der Reihenfolge der Wurfergebnisse abhängt[71], muss die Wahrscheinlichkeit der Stichprobe noch mit der Anzahl der möglichen Stichproben mit siebenmal Kopf und dreimal Zahl multipliziert werden (Bortz 1999: 99, Rost 2004: 304ff).[72]

Gl. (26) verdeutlicht die Idee der Maximum-Likelihood-Schätzung. Anstatt mit einer angenommenen Verteilung zu überprüfen, ob die Daten mit dem Modell vereinbar sind, geht die Maximum-Likelihood-Schätzung den umgekehrten Weg und sucht nach den Parametern der Verteilung, die die gegebenen Daten für das Modell maximal wahrscheinlich werden lassen (Mung 2003: 93).

Um nun den maximalen Wert, die maximale Wahrscheinlichkeit, von Gl. (26) zu bestimmen, muss deren Extremwert berechnet werden. Um den Prozess der Bildung der ersten Ableitung zu vereinfachen, wird der Ausdruck auf der rechten Seite von Gl. (26) logarithmiert. Denn dadurch wird das Produkt der Wahrscheinlichkeitsfunktionen zu einer Summe von Funktionen, die sich wesentlich einfacher ableiten lassen als das Produkt von Funktionen. Diese Transformation ist zulässig, da die logarithmierte Gleichung an der selben Stelle ein Maximum hat wie die unlogarithmierte Funktion.[73]

Es ergibt sich folgende Gleichung:

$$\begin{aligned} \ln p(x|\xi) &= \ln\binom{10}{7} + \ln \sum_{k=1}^{10} \xi^{x_k} \cdot (1-\xi)^{(1-x_k)} \\ &= \ln\binom{10}{7} + \ln \xi^7 + \ln(1-\xi)^3 \end{aligned} \quad . \qquad (27)$$

70 Das Ergebnis des *n*-ten Münzwurfs ist unabhängig vom Ergebnis des *(n-1)*-ten Münzwurfs. Die Münze hat kein Gedächtnis. Daher ergibt sich die Gesamtwahrscheinlichkeit als das Produkt der Einzelwahrscheinlichkeiten *n,* siehe z.B. Bortz 1999, Dietz, Kalof 2010, Wirtz, Nachtigall 2004, Tiede, Voss 2000.

71 Die Reihenfolge des Auftretens der Ereignisse ist hier kein Unterscheidungsmerkmal der Stichprobe. Z.B. wird das Wurfergebnis (Kopf, Kopf, Zahl, Zahl) identisch mit dem Wurfergebnis (Zahl, Kopf, Zahl, Kopf) angesehen.

72 Die Anzahl der möglichen Kombinationen aus einer Stichprobe vom Umfang *n* in der *k*-mal Kopf aufgetreten ist berechnet sich mit Hilfe des Binomialkoeffizienten: $\binom{n}{k} = \frac{n!}{k!\cdot(n-k)!}$ mit *n!* als der Fakultät der Zahl *n*. (Forster 1985: 3)

73 Die Logarithmusfunktion ist eine streng monotone Funktion und erhält daher die Lage des Extremwerts (Fischer 1974: 230, Rost 1994: 305, Forster 1985).

Gl. (27) ist die logarithmierte Likelihood Funktion des Beispiels.

Mit den üblichen Ableitungsregeln[74] ergibt sich als Ableitung von Gl. (27)

$$(\ln p(x|\xi))' = \frac{7}{\xi} + \frac{3}{(1-\xi)} \cdot (-1) \quad . \tag{28}$$

Die Nullstellen von Gl. (28) sind eine notwendige Bedingung für das Vorliegen eines Extremwerts der Funktion. Als Nullstelle von Gl. (28) ergibt sich $\xi=0,7$. Die zweite Ableitung von Gl. (27) an der Stelle $\xi=0,7$ ist kleiner Null: Für $\xi=0,7$ liegt also ein Maximum vor (Forster 1985: 109ff). Die Münze ist demnach nicht fair, sondern fällt, wie aus den Daten anschaulich ersichtlich ist, mit einer Wahrscheinlichkeit von $\xi=0,7$ auf den Kopf.

Für unser Modell ergibt sich daher folgende mittels des Maximum-Likelihood-Verfahren geschätzte Modell-/Verteilungsgleichung:

$$p(x|\xi) = \binom{10}{7} \prod_{k=1}^{10} 0,7^{x_k} \cdot (1-0,7)^{(1-x_k)} \quad .^{75} \tag{29}$$

Nun ist aber das vorgestellte Beispiel des Münzwurfs für komplexe Modelle aus folgenden Gründen untypisch:

1) Die Ableitung des Modells kann algebraisch, d.h. in geschlossener Form, angegeben werden.
2) Das Modell des Münzwurfs hat nur einen zu schätzenden Modellparameter.
3) Der einzige zu schätzende Parameter hängt nicht von einer zusätzlichen Modellannahme, wie z.B. einer Item-Response-Funktion ab.

Üblicherweise haben die betrachteten statistischen Modelle mehrere Unbekannte, wie z.B. im Fall von Personen, die Aufgaben lösen, und damit so viele Unbekannte wie sich als Summe j aus der Anzahl der beteiligten Personen und der verwendeten Aufgaben ergibt. Die Likelihood Funktion ist somit eine j-dimensionale Fläche im j-dimensionalen Parameterraum (Mung 2003: 93). In dieser j-dimensionalen Hyperfläche müssen die Maxima gesucht werden, die

74 Die Ableitung des Logarithmus $f(x)$ ist: $f'(x)=1/x$; Anwendung der Kettenregel (Forster 1985: 99ff)

75 Dieses Verfahren wurde erstmals von R.A. Fisher im Jahr 1925 vorgestellt.

man sich als die Berge (und Täler) innerhalb dieser *j*-dimensionalen Fläche vorstellen kann.

Um die Maxima einer solchen Likelihood Gleichung zu bestimmen, muss die Likelihood Gleichung des Modells nach jedem der zu bestimmenden Parameter partiell differenziert werden (siehe Forster 1986: 34ff). Es entsteht somit ein Gleichungssystem mit *j* Gleichungen und *j* Unbekannten, denn die Modellgleichung wurde *j*-mal partiell differenziert, jeweils nach jedem Personen- und jedem Aufgabenparameter. Durch jede partielle Differenzierung entsteht eine Gleichung in der im Extremfall alle Parameter als Unbekannte enthalten sein können.

Die Bedingung für das Vorliegen eines Maximums ist analog zu der Bedingung für Modelle mit nur einem Parameter: Die erste Ableitung muss an der Stelle des Extremwerts Null sein (Forster 1986: 35ff). Im Gegensatz zum eindimensionalen Fall ist durch die partielle Differenzierung nach allen Personen- und Aufgabenparametern jedoch ein Gleichungssystem entstanden, das aus allen ersten partiellen Ableitungen besteht. Daher muss nicht *eine* erste Ableitung an der Stelle des Extremwerts verschwinden, sondern im Fall einer *j*-dimensionalen Likelihood Funktionen müssen für das Vorliegen eines Extremwerts *alle j*-ten ersten partiellen Ableitungen an der Stelle des Extremwerts des *j*-dimensionalen Raums Null werden.[76] Es sind daher die Nullstellen des Gleichungssystems zu bestimmen, das aus den *j* partiellen Ableitungen entstanden ist.[77] Da es sich bei den Modellgleichungen grundsätzlich um nicht-lineare Gleichungen handelt, wie z.B. eine Exponentialfunktion bei der Gaußfunktion, das Verhältnis aus Exponentialfunktionen beim Rasch-Modell oder trigonometrische Funktionen, können die Lösungen solcher Systeme grundsätzlich nicht mehr in algebraischer Form angegeben werden. Zur Bestimmung der Nullstellen solcher Systeme werden daher numerische Verfahren zur Nullstellen Bestimmung angewendet, wie z.B. das Newton-Verfahren oder das Broyden-Verfahren, die iterativ, ausgehend von vorgegebenen Startwerten, eine Lösung des Gleichungssystems suchen.[78]

76 Wie im Falle von Funktionen mit nur einer Variablen ist dies nur eine notwendige Bedingung für die Existenz eines Extremwerts. Im eindimensionalen Fall muss anschließend noch das Verhalten der zweiten Ableitung an der Stelle des Extremwerts untersucht werden, im mehrdimensionalen Fall wird die Eigenschaft der Hesseschen Matrix an der Stelle des Extremwerts untersucht (Forster 1985: 112, Forster 1986: 59ff).

77 Analog zu dem aus der Schule bekannten Probleme, z.B. die drei Unbekannten eines linearen Gleichungssystems bestehend aus drei Gleichungen zu bestimmen: „drei Gleichungen mit drei Unbekannten“.

2.6.2 Konsistenz der Parameterschätzung der Maximum-Likelihood-Methode

Aufgrund der Komplexität des zu lösenden Gleichungssystems ergibt sich folgende Frage bzgl. der Lösungen des Maximum-Likelihood-Gleichungssystems, die von R.A. Fisher (1967) selbst aufgegriffen wurde:

Liefert die Methode der Maximum-Likelihood-Schätzung konsistente Schätzer der Parameter?

Dabei heißt ein Schätzwert konsistent, wenn bei zunehmendem Stichprobenumfang die Differenz aus wahrem und geschätztem Parameterwert immer geringer wird.

Formal ausgedrückt:
Ein Schätzer heißt konsistent, wenn für jedes $\varepsilon > 0$ und $\zeta > 0$ ein N existiert, s.d. die Wahrscheinlichkeit für

$$|Sch\ddot{a}tzwert - wahrem\ Wert\ des\ Parameters| < \varepsilon \tag{30}$$

größer als $1 - \zeta$ ist für alle $n > N$ ist, oder

$$p(|Sch\ddot{a}tzwert - Parameterwert| < \varepsilon) \to 1 \quad , \ f\ddot{u}r \ n \to \infty \quad , \tag{31}$$

mit p als der Wahrscheinlichkeit des Ereignisses (Kendall, Stuart 1951: 3, Bortz 1999: 96, Neyman, Scott 1948: 2, Dietz, Kalof 2009: 286, Rost 1994: 313).[79, 80]

Der Hinweis auf die Frage nach der Konsistenz der Maximum-Likelihood-Parameterschätzung weist auf ein auch für die Item-Response-Theorie zu lösendes Problem hin, dass 1948 von Neyman und Scott entdeckt und nach ihnen benannt wurde (siehe unten).[81]

78 Eine Beschreibung der Vorgehensweise findet sich in der Dokumentation der Computer Algebra Systemen von Mathematica® (Wolfram Research), hier der Befehl „FindRoot", und den Euler Math Toolbox, hier der Befehl „broyden" und im Anhang zur Beschreibung der Anwendung des trigonometrischen Verfahrens mit den Euler Math Toolbox.

79 Diese Definition beinhaltet eine offensichtliche Analogie zur mathematischen Definition der Konvergenz (Kendall, Stuart 1951: 3, Forster 1983: 19).

80 Im Vergleich zur Definition der Erwartungstreue einer Parameterschätzung ist die Forderung der Konsistenz der Schätzung eines Parameters eine schwächere Forderung, als die Forderung der Erwartungstreue/unbiasedness des Parameters (siehe Abschnitt 2.6.1). Daher folgt aus der Konsistenz eines Schätzers nicht die Erwartungstreue eines Schätzers (Kendall, Stuart 1951: 5).

81 Nach Fisher (1967) ist es immer möglich eine Parametrisierung der Daten zu finden, so dass die Parameter mittels Maximum-Likelihood konsistent geschätzt werden kön-

Zusammenfassend weisen Maximum-Likelihood-Schätzer für eine Vielzahl von Voraussetzungen folgende Eigenschaften auf (Kendall, Stuart 1951: 40):

1) Der oder die Maximum-Likelihood-Schätzer sind asymptotisch normal verteilt, d.h. mit wachsendem Stichprobenumfang nimmt die Verteilung der Maximum-Likelihood-Schätzer die Form einer Normalverteilung an.
2) Die Varianz der Maximum-Likelihood-Schätzer konvergiert mit wachsender Stichprobengröße gegen Null, d.h. der Fehler der Schätzung verschwindet (konsistente Schätzung)und die .
3) Die Maximum-Likelihoodlihood Schätzer sind effizient (Bortz 1999: 96, Dietz, Kalof 2009: 281), d.h. andere Schätzer der selben Parameter haben größere Varianzen, d.h. größere Fehler (Kendall, Stuart 1951: 40, Neyman, Scott 1948, Mung 2003: 96f, Dietz, Kalof 2009: 287, Fisher 1967: 14, Swaminathan 1983: 30).

2.6.3 Der Standardfehler des Maximum-Likelihood-Parameterschätzers

Parameter sind Determinanten des Datenhintergrunds und als solche stellen sie Aggregate dar, die den Datensatz nicht nur repräsentieren, sondern ihn auch rechnerisch – sozusagen von hinten nach vorne – generieren können. Die durch die Aggregation erzeugte Information in Form eines Parameter-Aggregatwertes kann nicht alleine an sich beurteilt werden, da die Relevanz der gewonnenen Information abhängig von dem "Fehler" dieser Aggregation ist. Die durch die Aggregation erzeugte Komplexitätsreduktion der Daten muss auf ihre Relevanz hin überprüft werden, z.B. durch weitere Parameter, mit deren Hilfe die Einschätzung des Parameters differenzierter erfolgen kann. Wie wichtig das sein kann, verdeutlicht die Interpretation der Durchschnittsnote einer Schulklasse (zur grundsätzlichen Problematik der Beurteilung von Schülerleistungen durch Zensuren siehe 1.4):

Angenommen, die Berechnung des arithmetischen Mittels der Noten einer Klassenarbeit einer Unterrichtsklasse im Unterrichtsfach Mathematik hätte den Parameterwert drei erzeugt. Dieser Parameterwert kann in Klassen entstehen, die über äußerst unterschiedliche Noten bzw. Leistungsverteilungen verfügen. So

nen (Fisher 1967: Kapitel 1 8).

könnte es eine Klasse geben, in der alle Schülerinnen und Schülern mit der Note drei beurteilt wurden und eine andere Klasse in der die eine Hälfte der Schülerinnen und Schüler die Note fünf und die andere Hälfte die Note eins erhalten haben.

Beide Klassen alleine mit Hilfe der arithmetisch gemittelten Durchschnittsnote zu vergleichen würde der Komplexitätsreduktion der Daten durch den Parameter nicht gerecht werden, da das Notenmuster, die Notenverteilung, der Klassen keine Berücksichtigung findet. Dem arithmetischen Mittel tritt daher üblicherweise die Standardabweichung[82] hinzu, die die Verteilung der Daten bzgl. des Mittelwerts abbildet.

Für Maximum-Likelihood-Parameter ist das Gütemaß für die Parameterschätzer[83] der Wert, den die von R.A. Fisher definierte Informationsfunktion für den Parameter annimmt. Die Informationsfunktion I für einen geschätzten Parameter ξ ergibt sich als Erwartungswert:

$$I(\xi) = -\overline{\left(\frac{\partial^2 \log L}{\partial \xi^2}\right)} \tag{32}$$

(Rost 2004: 357, Kendall, Stuart 1951: 10, Lord, Novick 1968: 419).

Es wird also der negative Erwartungswert der zweiten partiellen Ableitung der Likelihood Funktion des Modells nach dem Parameter ξ gebildet.

Die Standardschätzfehler, d.h. die Varianz des Parameterschätzfehlers E_{ξ}, berechnet sich aus Gl. (32) durch

$$s(E_{\xi}) = \frac{1}{\sqrt{I(\xi)}} \quad \text{(Rost 2004: 358, Fischer 1974: 293).} \tag{33}$$

Gl. (33) definiert den Standardschätzfehler eines Maximum-Likelihood-Parameterschätzers.

82 Die Standardabweichung ist definiert als die Wurzel aus der Summe der quadrierten Abweichung der einzelnen Daten vom Mittelwert (Varianz), die an der Anzahl der Datenpunkte relativiert wird. Die Standardabweichung erfüllt die Bedingung der kleinsten Quadrate (siehe z.B. Bortz 1999: 42).

83 Da innerhalb inferenzstatistischer Verfahren Parameter aus einer Stichprobe stellvertretend für die zu Grunde liegende Population berechnet werden, wird in diesem Zusammenhang auch von Parameterschätzung gesprochen: Schätzung des Parameters unter Verwendung einer repräsentativen Stichprobe. Parameterschätzwert wäre die genauere Bezeichnung.

Die Parameter bei den in dieser Arbeit vorgestellten Modellen werden durch das Maximum-Likelihood-Verfahren berechnet/geschätzt, siehe 2.4, 3.1.4, 3.1.5 und 4.8. Bei der Darstellung der Eigenschaften der Modelle wird daher nochmals auf das Problem des Fehlers der Parameterschätzung im jeweiligen Modell eingegangen.

2.6.4 Darstellung der Maximum-Likelihood-Methode für in einem Test zusammengefasste Items

Die eben gezeigte beispielhafte Berechnung des Maximum-Likelihood-Parameters für den Münzwurf wird nun verallgemeinert. Das Werfen immer derselben Münze kann als Synonym für das Lösen immer der gleichen Aufgabe angesehen werden, siehe 1.5. Ein Test besteht aber grundsätzlich nicht nur aus einer Aufgabe, sondern aus mehreren Aufgaben. Daher wird im Folgenden das Verteilungsmodell für mehrere dichotome Aufgaben entwickelt, deren Lösungswahrscheinlichkeiten zum einen von zwei Parametern und zum anderen von einer Item-Response-Funktion abhängen.

Das Verteilungsmodell für dichotome Variablen ist das so genannte Binomialmodell (Bortz 1999: 65ff, Harney 2003: 40f), siehe Gl. (25).

Da wie oben erwähnt in der Item-Response-Theorie die lokale stochastische Unabhängigkeit gilt oder, je nach Betrachtungsweise, auch die stochastische Unabhängigkeit, gilt für die Wahrscheinlichkeit der Lösung oder Nicht-Lösung von dichotomen Aufgaben die Berechnungsweise der Gesamtwahrscheinlichkeit für unabhängige Zufallsvariablen. D.h. die Wahrscheinlichkeiten werden multiplikativ verknüpft (Bortz 1999: 55, Tiede, Voss 2000: 30ff, Wirtz, Nachtigall 2004a: 23). Die Wahrscheinlichkeit des Antwortverhaltens einer Person bezüglich mehrerer Aufgaben ergibt sich daher aus der in Gl. (34) angegebenen „Verbundwahrscheinlichkeit“ für die betrachtete Person über alle Aufgaben:

$$p(x)=\prod_{l=1}^{N_i} q(x_l) \quad , \tag{34}$$

mit N_i als der Anzahl der von einer Person bearbeiteten Aufgaben und $q(x_l)$ als den Wahrscheinlichkeiten der Lösung der Aufgabe x_l. Gl. (34) stellt die Wahrscheinlichkeit des Antwortmusters einer Person dar.

Die Wahrscheinlichkeit ξ_l der Lösung einer Aufgabe x_l kann aber von Parametern abhängen. Im Falle der Item-Response-Theorie indizieren die Para-

meter, die die Lösungswahrscheinlichkeit einer Aufgabe beeinflussen, die Fähigkeitsausprägung der Person und die Schwierigkeit der Aufgaben. Dabei wird der Zusammenhang der Parameter und der Lösungswahrscheinlichkeit der Aufgaben durch eine zu wählende Item-Response-Funktion ϕ hergestellt. Daher muss Gl. (25) um diese parametrische Veränderung der Wahrscheinlichkeit ergänzt/vervollständigt werden:

$$q(x_l \mid \theta_k, \sigma_l) = \phi(\theta_k, \sigma_l)^{x_j} (1 - \phi(\theta_k, \sigma_l))^{1-x_j} \quad , \tag{35}$$

mit θ_k als der Ausprägung der Person k auf der betrachteten Fähigkeit/Dimension und σ_l als der Aufgabenschwierigkeit der Aufgabe x_l. Gl. (35) gibt somit die Wahrscheinlichkeitsverteilung der Aufgabe x_l in Abhängigkeit der Personenfähigkeit θ_k und der Aufgabenschwierigkeit σ_l an.

Gleichung (34) ändert sich dann analog:

$$p(x \mid \theta_k, \sigma_l) = \prod_{l=1}^{N_I} q(x_l \mid \theta_k, \sigma_l) \quad . \tag{36}$$

Gl. (36) gilt für alle Personen k. Daher gelangt man zu dem Wahrscheinlichkeitsmodell der Gesamtheit der Daten, indem Gl. (36) auf alle Personen erweitert wird.

$$p(x \mid \theta, \sigma) = \prod_{k=1}^{N_P} \prod_{l=1}^{N_I} q(x_l \mid \theta_k, \sigma_l) \quad , \tag{37}$$

mit N_P als der Anzahl der Personen und N_I als der Anzahl der zu lösenden Aufgaben.

Man erhält so das Wahrscheinlichkeitsmodell aller betrachteten Personen und Aufgaben. Gl. (37) stellt die Likelihood Gleichung (siehe Abschnitt 2.6.2) der Item-Response-Theorie dar.

Wie im vorangegangenen Kapitel beschrieben soll die Gesamtwahrscheinlichkeit des Modells in Abhängigkeit der Parameter Aufgabenschwierigkeit σ_l und Personenfähigkeit θ_k bei zu wählender Item-Response-Funktion und gegebenen Daten maximiert werden. Dazu findet das beschriebene Maximum-Likelihood-Verfahren (siehe Abschnitt 2.6.2) Anwendung.

Gl. (37) trifft jedoch noch keine Aussage darüber in welcher funktionalen Beziehung/Verbindung[84] die Parameter Aufgabenschwierigkeit σ_l und Personenfähigkeit θ_k zur Wahrscheinlichkeit der Lösung einer Aufgabe l stehen. Dieser Zusammenhang wird durch die eben eingeführte zu wählende Item-Response-Funktion ϕ hergestellt. Es ergibt sich somit aus Gl. (37):

$$p(x|\phi(\theta,\sigma)) = \prod_{k=1}^{N_P} \prod_{l=1}^{N_I} q(x_l|\phi(\theta_k,\sigma_l)) \quad . \tag{38}$$

Nach Bildung der partiellen Ableitungen nach allen Parametern ergibt sich für das Binomialmodell folgendes allgemeines System von Maximum-Likelihood-Gleichungen:

$$\begin{aligned} 0 &= \sum_{l=1}^{N_I} \frac{\partial \phi_{kl}}{\partial \theta_k} \cdot \left[\frac{x_{kl}}{\phi_{kl}} - \frac{1-x_{kl}}{1-\phi_{kl}} \right], \qquad k = 1, \ldots, N_P \\ 0 &= \sum_{k=1}^{N_P} \frac{\partial \phi_{kl}}{\partial \sigma_l} \cdot \left[\frac{x_{kl}}{\phi_{kl}} - \frac{1-x_{kl}}{1-\phi_{kl}} \right], \qquad l = 1, \ldots, (N_I - 1) \end{aligned} \tag{39}$$

(Harney, Harney, Fuhrmann 2006: 45).

Das System der Maximum-Likelihood-Gleichungen (39) zur Bestimmung der Nullstellen und damit der Parameter ist überbestimmt.[85] Dies bedeutet, dass ein zu bestimmender Parameter frei gewählt werden muss, um die anderen Parameter bestimmen zu können. Zum Beispiel könnte der Aufgabe mit dem höchsten Aufgabenscore der Wert Null zugewiesen werden. Üblicherweise wird jedoch eine Summennormierung der Aufgabenparameter vorgenommen, d.h. dass die Summe über alle Aufgabenparameter Null ergeben soll $\sum_{l=1}^{N_I} \sigma_l = 0$. So erhalten die Aufgaben, die sich genau in der Mitte der Schwierigkeitsskala befinden den Schwierigkeitsparameter Null.

Die Wahl der Item-Response-Funktion ϕ ist an dieser Stelle nicht festgelegt – bis auf die Bedingung, dass ihr Wertebereich dem Wahrscheinlichkeitsraum entsprechen und eine konsistente Schätzung ermöglichen muss.

84 Die Item-Response-Funktion wird daher auch oft Link-Funktion genannt (vgl. u.a. de Boeck, Wilson 2004: 30, Skrondal, Rabe-Hesketh 2004: 24).

85 Dies erkennt man an der Indexgrenze für die Aufgabengleichungen des Systems, zweite Zeile Gl. (39).

Dabei sind neben den zu Anfang vorgestellten sigmoiden Item-Response-Funktionen auch alternative Funktionsverläufe vorstellbar, wie z.B.:

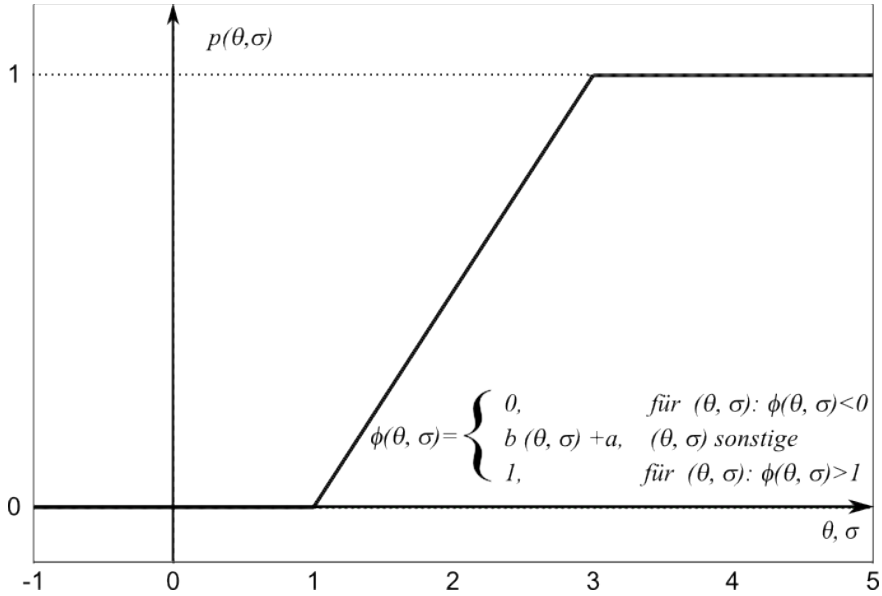


Abbildung 14: Beispiel einer quasi linearen Item-Response-Funktion

oder

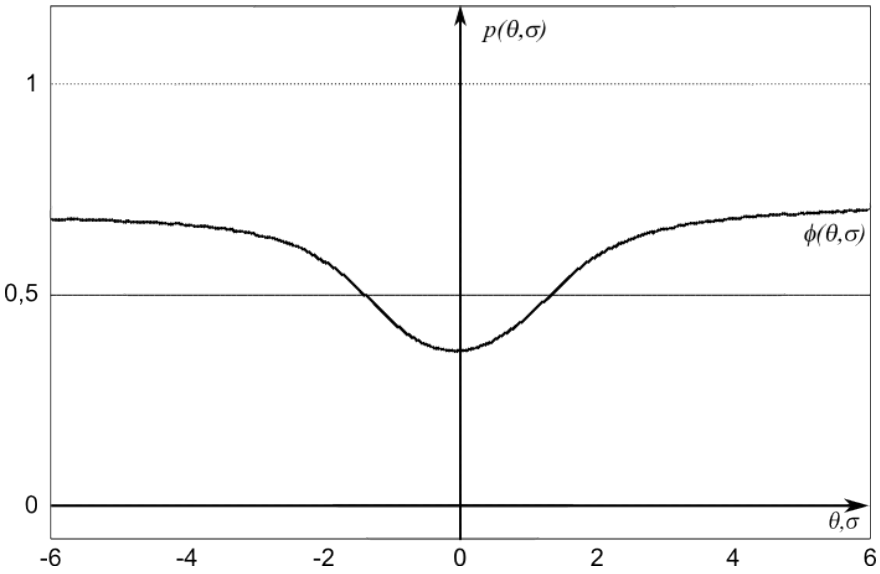


Abbildung 15: U-Förmige Item-Response-Funktion

Für den Verlauf in Abbildung 14 wäre zu begründen, weshalb die Item-Response-Funktion ab einem bestimmten Verhältnis von σ und θ nur noch den konstanten Wert Eins oder Null annimmt[86] (Rost 2004: 109, van der Linden, Hambleton 1991: 5). Abbildung 15 ist ein Verlauf, der aus Wachstumsprozessen bekannt ist, wie z.B. die Zu- und Abnahme der körperlichen Kraft im Laufe des Lebensalters eines Menschen oder der Sterberate (Mortalität).[87]

86 Wilson (1989: 278ff) betrachtet dazu das Saltusmodell, in dem die Wahrscheinlichkeiten sprunghaft ansteigen, was inhaltlich analog zur Begrenzung der Funktionsverlaufs wie in Abbildung 15 zu interpretieren ist bzw. begründet werden muss..

87 Sterberate/Sterbewahrscheinlichkeit: „Wahrscheinlichkeit, dass ein x-jähriger Mensch im Laufe des (x+1)-ten Lebensjahres stirbt“ (Scheid 1985: 62). Siehe auch beispielhaft Bayerisches Landesamt für Statistik (2015).

2.6.5 Maximum-Likelihood-Schätzungen für Modelle mit „incidental“ und „structural“ Parametern

Neyman und Scott (1948) zeigen, dass für bestimmte Modellstrukturen das Maximum-Likelihood-Verfahren keine konsistente Schätzung der Parameter zulässt. Für Neyman und Scott werden diese Modellstrukturen durch die mögliche Unterteilung der Modellparameter in die von ihnen eingeführte Unterscheidung der Parameter in „incidental“ und „structural“ Parameter gekennzeichnet. Item-Response-Modelle erfüllen diese von Neyman und Scott aufgestellten Strukturvoraussetzungen.

Neyman und Scott (1948) beschreiben in ihrem Artikel das Phänomen, dass für bestimmte statistische Modelle die ermittelten Maximum-Likelihood-Schätzer nicht zwingend konsistent sind. D.h.: Mit zunehmendem Stichprobenumfang konvergieren die Ergebnisse der Berechnung nicht gegen den wahren Wert (siehe die Definition der Konsistenz in Abschnitt 2.6.2).

Die von diesem Phänomen betroffenen statistischen Modelle haben dabei folgende Form:

1) Durch die betrachteten Zufallsvariablen lassen sich trivialerweise nicht nur unendliche viele Datenpunkte erzeugen, sondern gleichzeitig steigt mit Zunahme weiterer Datenpunkte auch die Anzahl der zu schätzenden Parameter des Modells an. Deren Anzahl dann ebenfalls unendlich werden kann.
2) Die betrachteten Zufallsvariablen können in Sequenzen von Beobachtungen aufgeteilt werden, wie z.B. Messungen an stellaren Objekten oder Messungen zu ein und demselben Phänomen in unterschiedlichen Laboratorien, so dass die zu ermittelnden Parameter in „structural” und „incidental” Parameter unterteilt werden können. Dies bedeutet, dass eine begrenzte Anzahl der Parameter („structural“) in jeder dieser Sequenzen der so aufgeteilten Zufallsvariablen erscheint, eine unbegrenzte Anzahl („incidental“) der Parameter aber nur in einer beschränkten Anzahl von Sequenzen.

Beispielhaft für ein solches Modell sind die Messungen verschiedener Phänomene an einer Sonne von verschiedenen Labors aus. Die „structural“ Parameter sind in diesem Fall die zu beobachtenden Phänomene der Sonne und die „incidental“ Parameter die Messsequenzen der einzelnen Labors mit ihren „spezifischen messfehlererzeugenden Eigenheiten”. Oder man betrachtet die Messung

einer astronomischen Konstanten anhand verschiedener Sonnen. Die Sequenzen entsprechen den wiederholten Messungen der astronomischen Konstante an den verschiedenen Sonnen. Der „structural“ Parameter ist gleichbedeutend mit der astronomischen Konstante.

Für die betrachteten Modelle können die „incidental“ Parameter nicht konsistent geschätzt werden, da sie nur in einer finiten Anzahl von Sequenzen vorkommen. Aber das essenzielle Problem ist, dass auch die „structural“ Parameter grundsätzlich nicht konsistent geschätzt werden können: Selbst dann nicht, wenn die Anzahl der Sequenzen unendlich groß wird. Durch die Hinzunahme weiterer Sequenzen wird durch jede hinzu genommene Sequenz mindestens ein weiterer zu schätzender Parameter dem Modell hinzugefügt. Für das Beispiel des Labors gilt: Jedes hinzu genommene Labor tritt mit seinem „typischen“ Messfehler dem Modell hinzu.

Die Parameterschätzung einer jeden Sequenz selbst lässt erst dann eine konsistente Schätzung zu, wenn auch die Länge der Sequenz im betrachteten Modell beliebig vergrößert werden kann.

Bezogen auf die Item-Response-Theorie ergibt sich in der Terminologie von Neyman und Scott nunmehr folgende Gegenüberstellung:

1) Geht man von einer Zufallsziehung der Personen aus, dann sind die Aufgaben die „structural“ und die Personen die „incidental“ Parameter.
2) Interpretiert man die Aufgaben als „ausgeloste“ Aufgaben aus einem Itemuniversum, dann sind die Aufgaben die „incidental“ und die Personen die „structural“ Parameter.

Die nachfolgende Argumentation geht von der ersten Betrachtungsweise aus, d.h. die Sequenz ist die Abfolge von Aufgaben, die eine Person löst, also der Test. Die Anzahl der Sequenzen kann durch Vergrößerung der Personenstichprobe „beliebig” vergrößert werden, wobei jede Person immer nur in einer Sequenz vorkommt, da jede Person nur einen Test bearbeitet. Jedoch muss für jede weiter hinzu genommene Person auch ein entsprechender Personenparameter mit in die Schätzung aufgenommen werden.

Der Test selbst lässt sich nicht beliebig verlängern, da die Bearbeitungszeit nicht beliebig gesteigert werden kann und zudem die Bearbeitungsfähigkeit und Motivation der zu testenden Personen durch Ermüdung mit der Zeit abnimmt (Liu, Harris, Schmidt 2007: 1064, Newman, Just 2005: 90).

Nach Neyman und Scott ist somit die Schätzung der Aufgabenparameter als den „structural“ Parametern selbst bei Vergrößerung der Personenstichprobe nicht konsistent möglich.

Als Beispiel der nicht konsistenten Schätzung im Falle von „structural“ und „incidental“ Parametern verwenden Neyman und Scott die Schätzung des Standardfehlers einer Sequenz von Messungen bei konstantem Messinstrument an unterschiedlichen Messobjekten, wie z.B. die Messung von Personenfähigkeiten mittels Aufgaben eines Tests als einem konstanten Messinstrument.

Auf dieser Grundlage berechnen Neyman und Scott den Maximum-Likelihood-Schätzer der Standardabweichung der Gaußverteilung. Der Schätzer wird als konstant für jede Sequenz angesehen, da das Messinstrument unabhängig von dem zu messenden und für jede Sequenz unterschiedlichen Test-Mittelwert, immer identisch exakt misst.

Die zu schätzende Varianz ergibt sich dann zu

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{S} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{\sum_{i=1}^{S} n_i} \tag{40}$$

(S = Anzahl der Sequenzen, n_i = Sequenzlänge bzw. Testlänge, x_{ij} = der Messwerte, $\bar{x}_i$ = Mittelwert der Sequenz/Score).

Für $n_i = n$ für alle S ergibt sich

$$\hat{\sigma}^2 = \frac{1}{S} \sum_{i=1}^{S} S_i^2,$$

$$S_i^2 = \frac{1}{n} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2. \tag{41}$$

Der Erwartungswert für jede einzelne Sequenz S_i^2 ergibt sich zu

$$E(S_i^2) = \frac{n-1}{n} \sigma^2 \tag{42}$$

(Bortz 1999: 91, Rohwer, Pötter 20002: 293, Dietz, Kalof 2009: 266ff, Kendall, Stuart 1947: 228ff) mit σ^2 als der „wahren“ Varianz.

Der Erwartungswert von Gl. (42) ist für alle Sequenzen S_i gleich und ergibt sich demnach S-mal. Zusammen mit der ersten Zeile von Gl. (41) folgt daraus

$$\lim_{S\to\infty} \hat{\sigma}^2 = \frac{n}{n-1}\sigma^2 \tag{43}$$

Man erkennt, dass diese Schätzung nur von der Länge der verwendeten Sequenz abhängig ist und nicht von der Anzahl der Sequenzen. Selbst wenn die Anzahl der Sequenzen gegen Unendlich anwächst, konvergiert der Fehler der Messung nicht gegen Null. Die Messung bleibt in Abhängigkeit der Testlänge n konstant ungenau (vgl. Neyman, Scott 1948, Andersen 1973b, Zellner 1996, Spanos 2013, Baker, Seock-Ho 2004, Fischer 1974: 350, Lancaster 2000, Swaminathan 1983, Glas 1988, Bock 1997, Allison 2009, Bock, Moustaki 2007, Tuerlinckx, Rijmen 2004).

Eine Illustration der inkonsistenten Schätzung der Aufgabenparameter im Rasch-Modell für zwei Aufgaben findet sich bei Fischer (1974: 259ff), und zwar basierend auf einem Aufsatz von Andersen über „Conditional Inference For Multiple-Choice Questionaries" (1973). Es dauerte 20 Jahre, bis eine Verallgemeinerung des „2-Aufgaben" Beispiels bei Andersen auf beliebig viele Items gelang und somit die inkonsistente Schätzung des Rasch-Modells bei Anwendung der Maximum-Likelihood-Methode allgemein bewiesen werden konnte (Ghosh 1994).

An verschiedenen Stellen der Arbeit wird gezeigt, wie das Rasch-Modell und wie das trigonometrische Verfahren das Neyman und Scott Problem bearbeiten – siehe 3.1.4 und 4.10.

Die Lösung des Neyman und Scott Problems ist essentiell für die Berechnung relevanter Parameter in einem Modell mit „structural" und „incidental" Parametern, denn ohne das Wissen um die Möglichkeit einer konsistenten Schätzung der Parameter bliebe ein Modell quantitativ gehaltlos.

http://www.springer.com/978-3-658-19240-2

Die trigonometrische Parametrisierung von Kompetenzen
Zur Methodologie der probabilistischen Bildungsforschung
Fuhrmann, C.
2018, XXVII, 358 S. 60 Abb., Softcover
ISBN: 978-3-658-19240-2