22. Tagung Visselhövede 2010
Rainer Oestereich

Kleine Kritik des üblichen Vorgehens bei der Skalenentwicklung und Hinweis auf eine praktikable Alternative: Das Rasch Modell

1. Das übliche Vorgehen
2. 2 Fragen
2a. Skalenniveau?
2b. Summenwerte?
3. Die Raschskalierung

- Bitte kreuzen Sie an, wie stark folgende Feststellungen auf Sie zutreffen.
- Lassen Sie bitte keinen Satz aus.

	trifft genau zu	trifft ziemlich zu	trifft etwas zu	trifft nicht zu
1. Mich haben einige Dinge besonders angeregt oder interessiert.	4	3	2	1
2. Es lief bei mir gut.	4	3	2	1
3. Ich habe mich gefreut, weil mein Leben in Ordnung war.	4	3	2	1
4. Mein Leben hat mir Freude gemacht.	4	3	2	1
5. Ich habe schöne Sachen erlebt.	4	3	2	1
	trifft genau zu	trifft ziemlich zu	trifft etwas zu	trifft nicht zu

- Bitte kreuzen Sie an, wie stark folgende Feststellungen auf Sie zutreffen.
- Lassen Sie bitte keinen Satz aus.

	trifft genau zu	trifft ziemlich zu	trifft etwas zu	trifft nicht zu	
1. Mich haben einige Dinge besonders angeregt oder interessiert.	4	⊗3	2	1	3
2. Es lief bei mir gut.	4	3	⊗2	1	2
3. Ich habe mich gefreut, weil mein Leben in Ordnung war.	⊗4	3	2	1	4
4. Mein Leben hat mir Freude gemacht.	4	3	2	⊗1	1
5. Ich habe schöne Sachen erlebt.	4	⊗3	2	1	3
	trifft genau zu	trifft ziemlich zu	trifft etwas zu	trifft nicht zu	

∑13

Skala „Lebenszufriedenheit“ - Ungeeignetheit eines Items

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
LEBZ1	**7,6637**	**7,6927**	**,4466**	**,8844**
LEBZ2	7,5919	7,3417	,7098	,8141
LEBZ3	7,5381	6,9163	,7285	,8066
LEBZ4	7,6906	7,1426	,7361	,8065
LEBZ5	7,7399	6,6348	,7557	,7983

Reliability Coefficients

N of Cases = 223,0 N of Items = 5

Alpha = ,8535

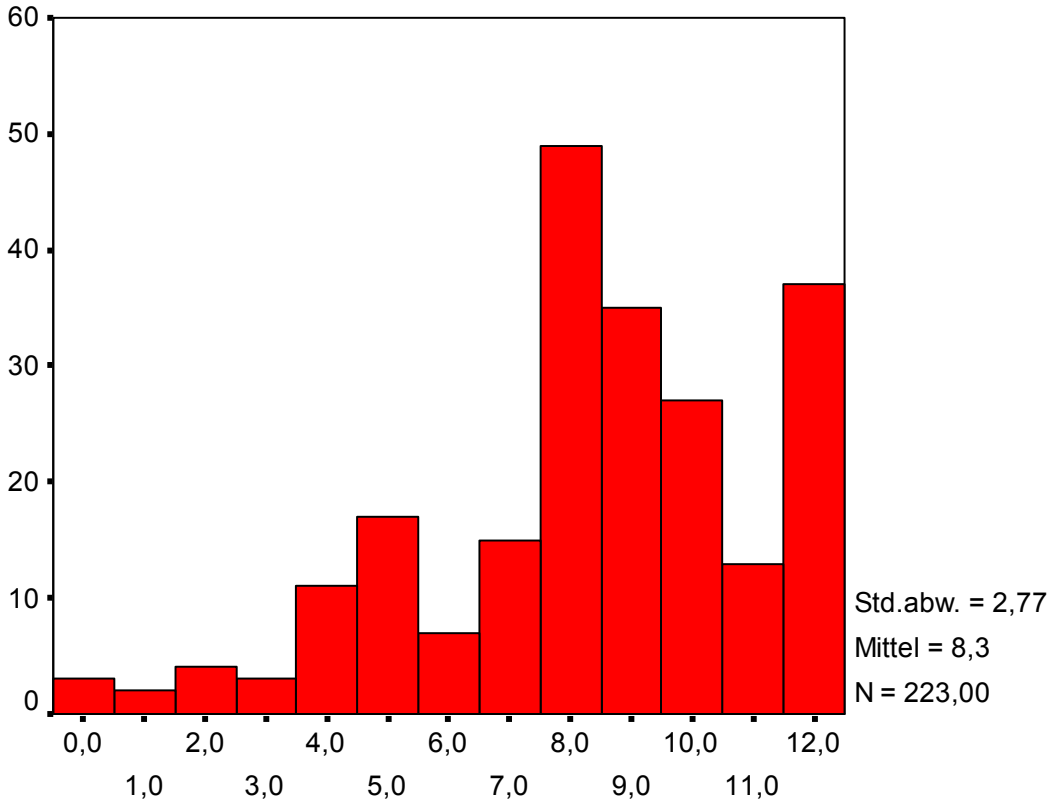
Skala „Lebenszufriedenheit“ - Ungeeignetheit eines Items

Im folgenden geht es darum,
wie Sie sich in den letzten 12 Monaten gefühlt haben

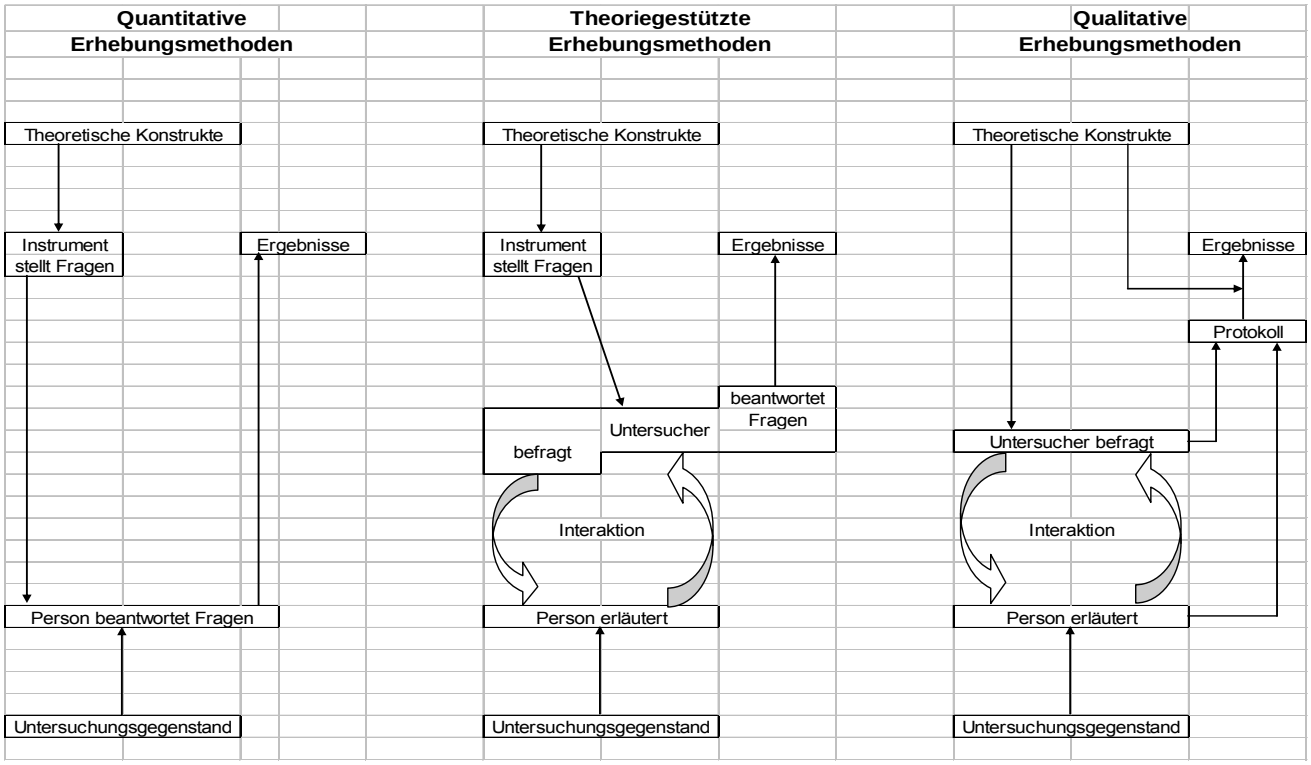
1. ~~Mich haben einige Dinge besonders angeregt oder interessiert.~~
2. Es lief bei mir gut.
3. Ich habe mich gefreut, weil mein Leben in Ordnung war.
4. Mein Leben hat mir Freude gemacht.
5. Ich habe schöne Sachen erlebt.

trifft genau zu	trifft ziemlich zu	trifft etwas zu	trifft nicht zu
3	**2**	**1**	**0**

Skala „Lebenszufriedenheit“



LEBENZ



Grundannahmen der klassischen diagnostischen Testtheorie und der Begriff „Reliabilität“

Grundannahmen ("Axiome") der klassischen Testtheorie

1. *X = Tau + E*

Ein **Messwert x** ergibt sich aus dem **true-score "tau"** plus einem **Messfehler "e"**

2. *Erw(E)* = 0

Der Erwartungswert für Messfehler ist stets null

3. *r(Tau,E)* = 0

Die Korrelation (bzw. Kovarianz) von true-score und Messfehler ist null (z.B.: bei höherem true-score wird nicht auch ein höherer Messfehler erwartet)

4. *r(E,E')* = 0

Die Korrelation (bzw. Kovarianz) zwischen Messfehlern verschiedener Tests ist null (z.B.: gab es bei einer Messung einen hohen Messfehler, impliziert dies nichts über die Messfehler anderer Messungen)

Gemessen werden können jedoch nur die Werte *x* (*Tau* und *E* sind unbekannt).

Reliabilität eines Tests:

Reliabilität = $\frac{\sigma^2(Tau)}{\sigma^2(Tau)+\sigma^2(E)} = \frac{\sigma^2(Tau)}{\sigma^2(X)}$ = "wahre Varianz"/Gesamtvarianz

Messfehler-Varianz: $\sigma_E^2 = \sigma_X^2(1-rel)$

Wie läßt sich die Reliabilität bestimmen?

4. Radikale Fortsetzung der Testhalbierung und "interne Konsistenz"

Wie kann ein Test "halbiert" werden?

Ist z.B. die Item-Anzahl gerade, gibt es
1/2 mal (2n über n) mögliche Test-Teilungen.

Werden nun sämtliche möglichen Test-Teil-Paare eines Tests betrachtet und wird die Paralleltest-Reliabilität (α, s.o.) jedes dieser Paare ermittelt und werden diese Reliabilitäten arithmetisch gemittelt ergibt sich ein Wert, für den sich zeigen läßt, dass er der Reliabilität des Gesamt-Tests enspricht (vorausgesetzt die Test-Teil-Paare sind essentiell Tau-Äquivalent).

Cronbach hat gezeigt, dass eine solche Berechnung mit dem Ergebnis folgender Formel identisch ist:

Cronbachs' α = $\frac{p}{1-p}\left(1-\frac{\sum_{i=1}^{p} s^2(Item\ i)}{s^2(Testwert\ X)}\right)$ wobei p = Anzahl der Items

Bei gleicher Varianz der Items ergibt sich Cronbachs Alpha als:

$\alpha = \frac{p\,\bar{r}}{1+(p-1)\,\bar{r}}$

wobei p = Anzahl der Items und $\bar{r}$ = durchschnittliche Korrelation

Cronbach´s Alpha ist eine Größe zwischen 0,0 und 1,0 und ist wie folgt bestimmt:

$$\alpha = \frac{K\,\bar{r}}{1+(K-1)\,\bar{r}}$$

wobei
K : Anzahl der Items
$\bar{r}$: durchschnittliche Korrelation zwischen den Items

Generell gilt:

Alpha unter 0,7: Skala nicht oder kaum brauchbar
Alpha 0,7 bis 0,75: Skala ausreichend
Alpha 0,75 bis 0,8: Skala befriedigend
Alpha ab 0,8: Skala gut (ab 0,9 sehr gut)

2. Zwei Fragen

2a. Haben die (Personen-)Scores
Intervall-Skalen-Charakter?

2b. Da die Scores sich als Summenwerte
(oder Derivate davon) ergeben,
sind die Summenwerte
„suffiziente“ Statistiken?

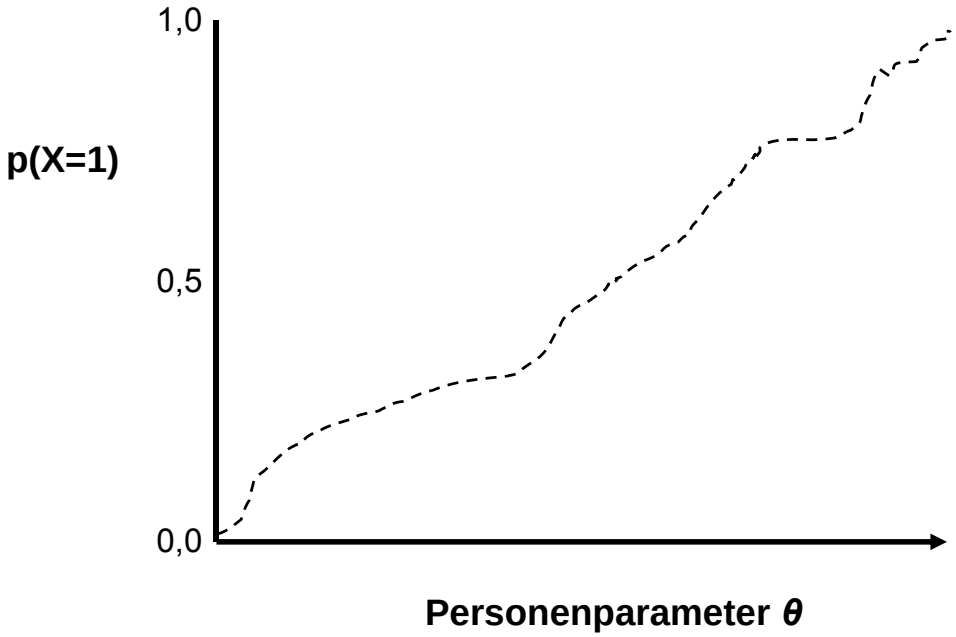
Skalen-Niveaus	definierte Relationen im empirischen Relativ	definierte Relationen im numerischen Relativ	zulässige Trans-formationen im numerischen Relativ	Beispiele	Anmerkungen
Verhältnis-Skala (Absolut-Skala) (Kardinal-Skala)	zusätzlich Größen-verhältnisse	$Y = a \cdot X$	alle linearen Transformationen ohne Nullpunkt-verschiebung $y_i = a \cdot x_i$	Länge, Gewicht, Zeitdauer Kelvin-Skala Häufigkeiten !!	absoluter bzw. „natürlicher“ Nullpunkt (gibt es ihn real?)
Intervall-Skala (Kardinal-Skala)	zusätzlich Abstände	$\|Y-X\| = \|U-V\|$	alle linearen Transformationen $y_i = a \cdot x_i + b$	Celsius-Skala Jahreszahl Uhrzeit?	positive und zugleich negative Werte möglich
Ordinal-Skala (Rang-Skala)	zusätzlich größer/kleiner	$Y > X$	alle monotonen Transformationen (z.B. Potenzieren mit 2, 3 usw.)	Schul-/Prüfungs-note Ausbildungs-niveau Rangliste	vollständige oder geteilte Ränge ?
Nominal-Skala (Kategorial-Skala)	Gleichheit/ Ungleichheit	$Y = X$ $Y \neq X$	alle Tranformationen, die Ungleichheit erhalten (nicht z.B. Potenzieren mit 0)	Frau/Mann Studienrichtung bevorzugte Partei	Problem der Eindeutigkeit der kategorialen Zuordnung

2. Zwei Fragen

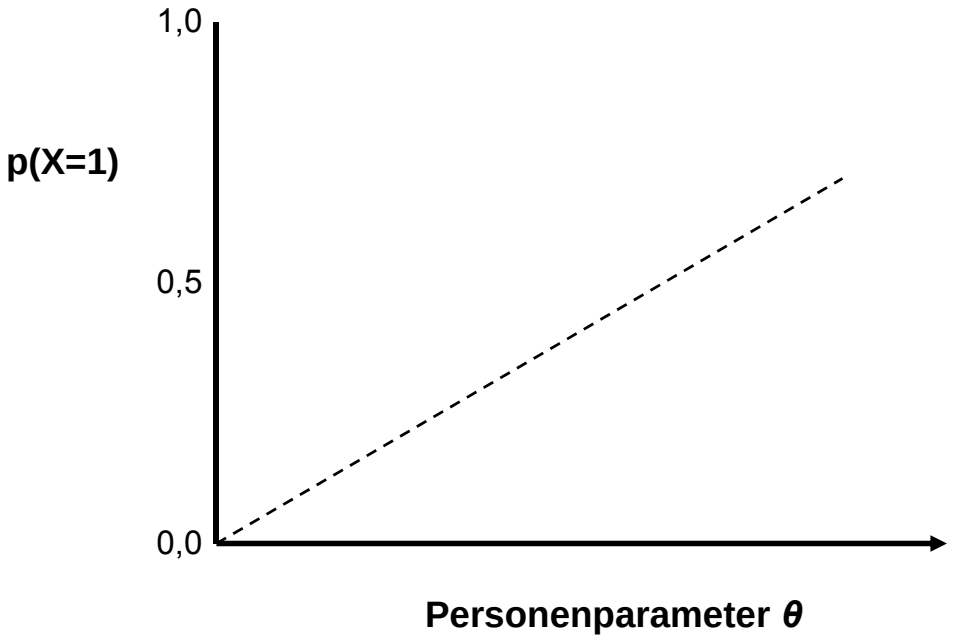
2a. Haben die (Personen-)Scores Intervall-Skalen-Charakter? Nein

2b. Da die Scores sich als Summenwerte (oder Derivate davon) ergeben, sind die Summenwerte „suffiziente“ Statistiken?

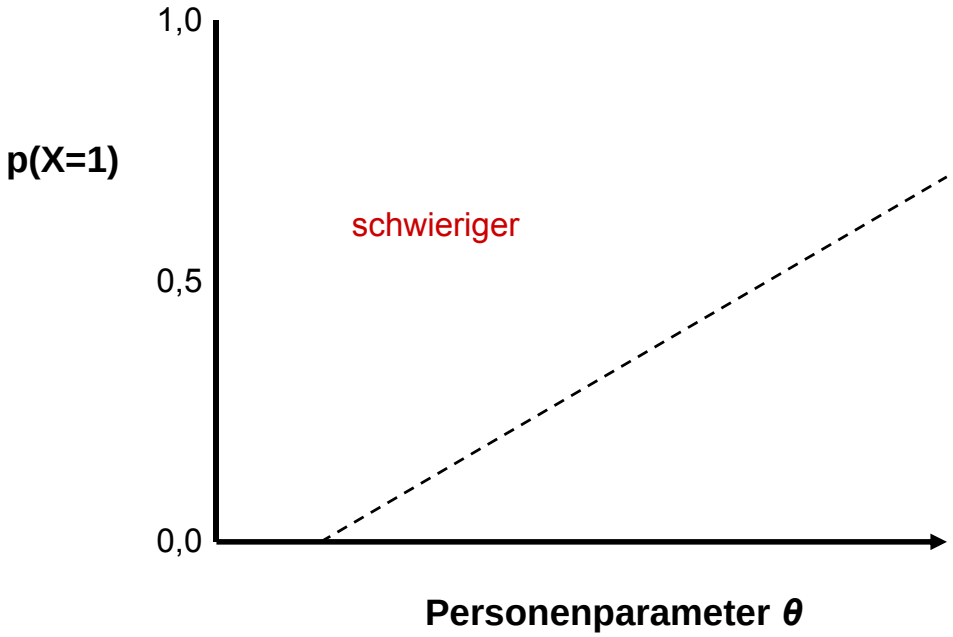
ICC


1,0
p(X=1)
0,5
0,0
Personenparameter θ

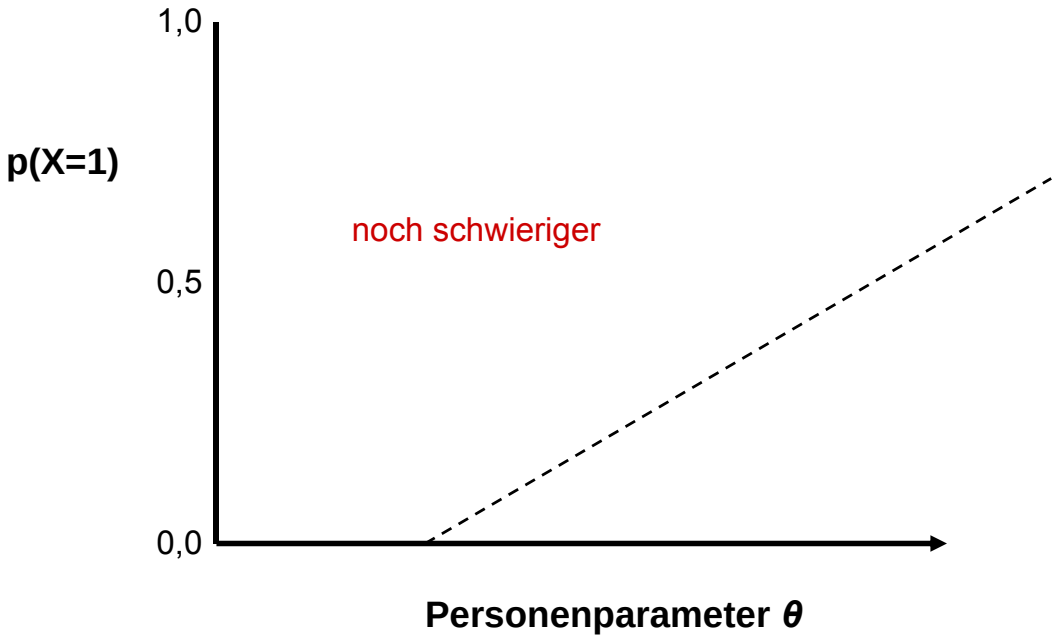
ICC


1,0
p(X=1)
0,5
0,0
Personenparameter θ

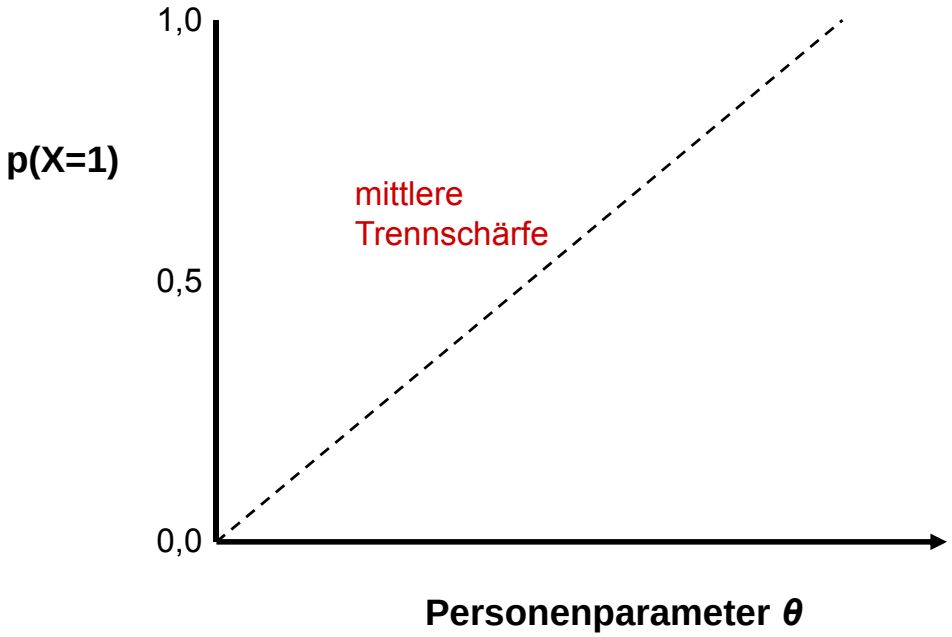
ICC


1,0
p(X=1)
schwieriger
0,5
0,0
Personenparameter θ

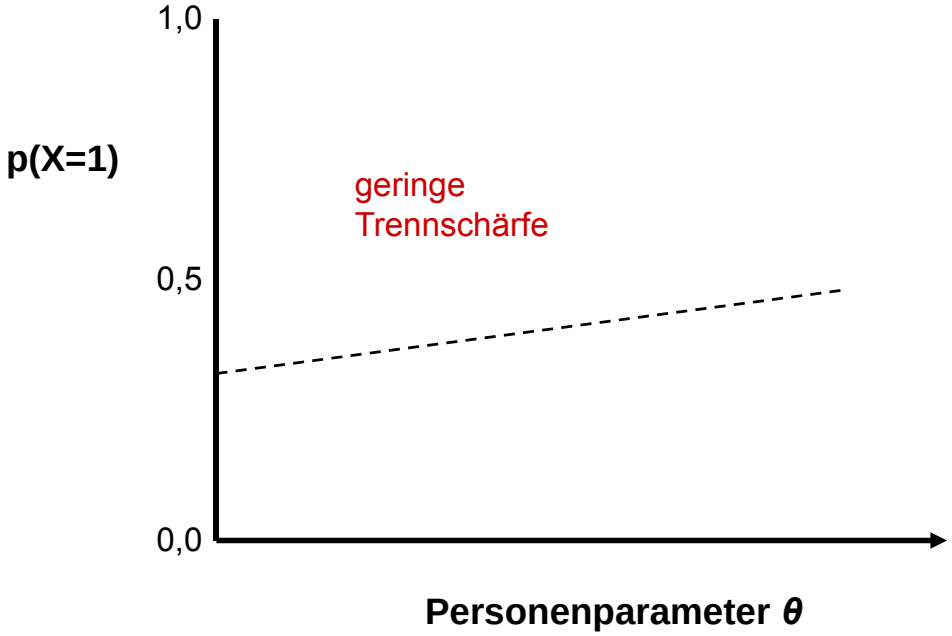
ICC


1,0
p(X=1)
noch schwieriger
0,5
0,0
Personenparameter θ

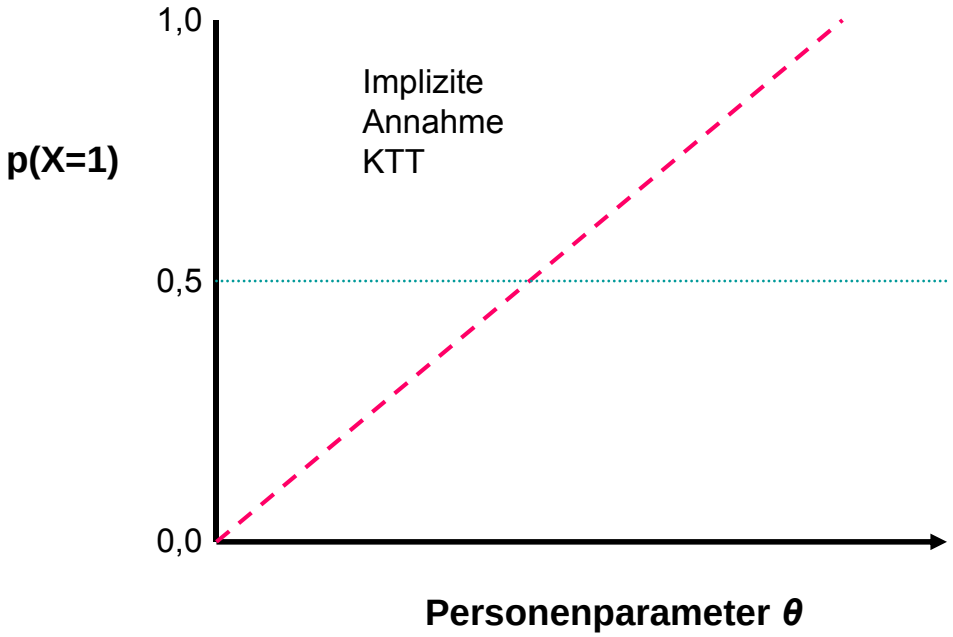
ICC


1,0
p(X=1)
mittlere
Trennschärfe
0,5
0,0
Personenparameter θ

ICC


1,0
p(X=1)
geringe
Trennschärfe
0,5
0,0
Personenparameter θ

ICC

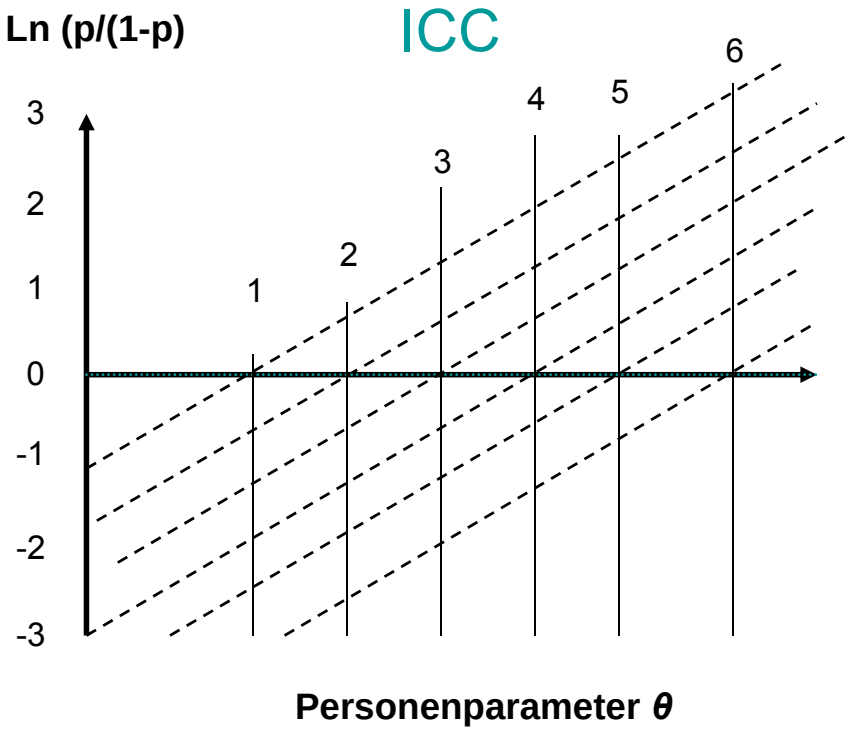


2. Zwei Fragen

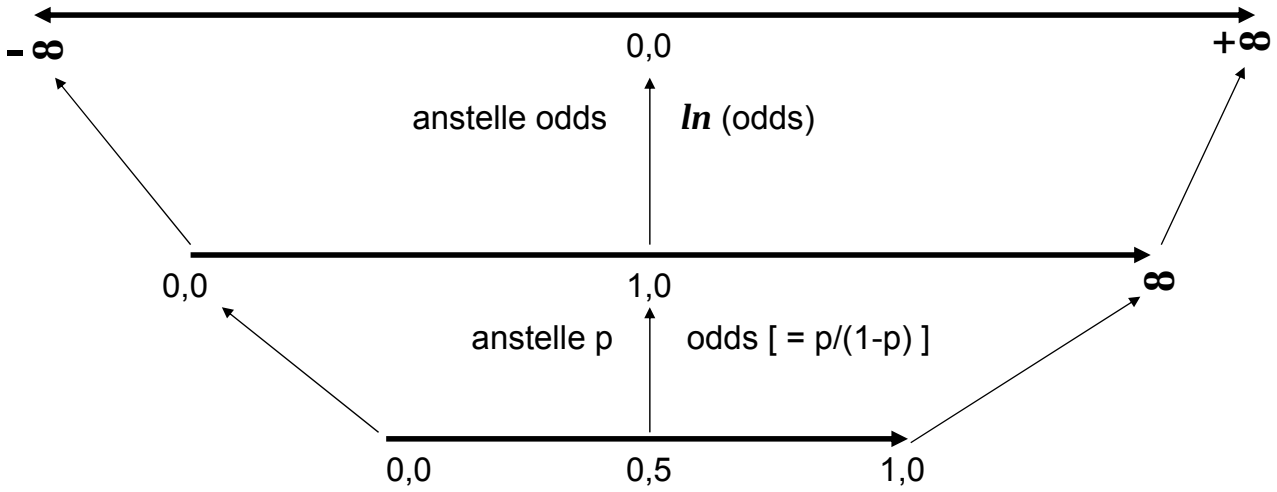
2a. Haben die (Personen-)Scores Intervall-Skalen-Charakter? Nein

2b. Da die Scores sich als Summenwerte (oder Derivate davon) ergeben, sind die Summenwerte „suffiziente“ Statistiken? Nein

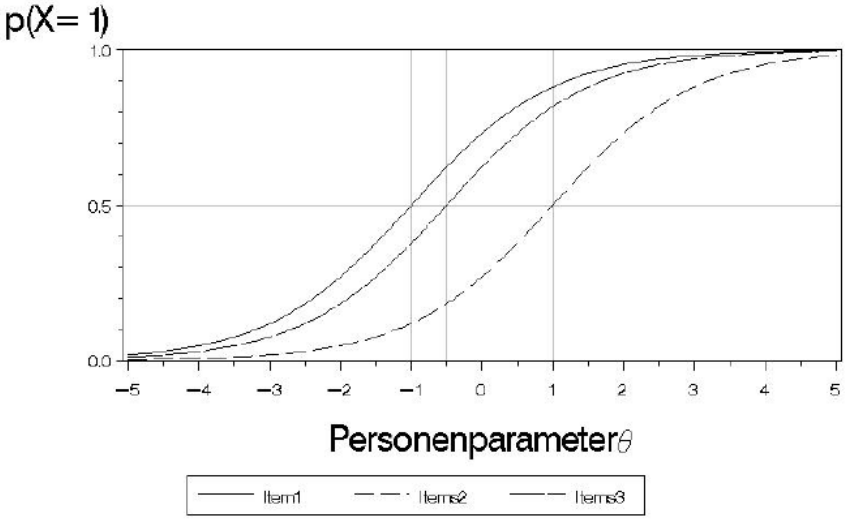
3 Die Raschskalierung

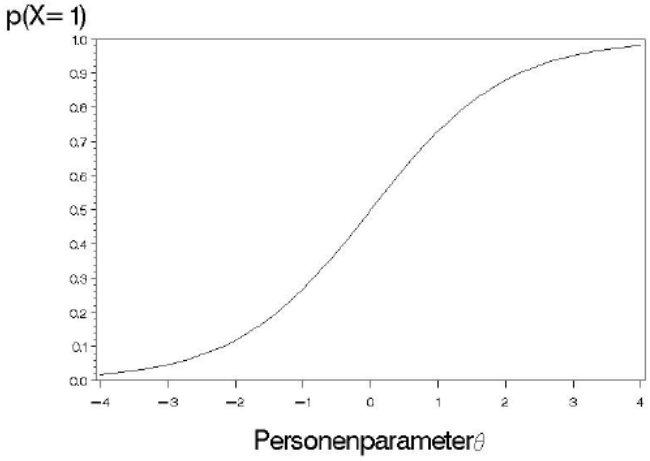


Personenparameter θ



p	p=p/(1-p)	ln(p/(1-p)
0,0001	0,00	-9,21
0,001	0,00	-6,91
0,05	0,05	-2,94
0,1	0,11	-2,20
0,2	0,25	-1,39
0,4	0,67	-0,41
0,5	1	0
0,6	1,5	0,41
0,8	4	1,39
0,9	9	2,20
0,95	19	2,94
0,999	999	6,91
0,9999	9999	9,21


p(X= 1)
1.0
0.5
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Personenparameterθ
Item1
Items2
Items3


p(X= 1)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−4
−3
−2
−1
0
1
2
3
4
Personenparameterθ

Nicht mehr bestehende Probleme der Anwendung der Rasch-Skalierung

(a) Soll man die logarithmierten oder ent-logarithmierten Personenparameter in den späteren Datenauswertungen verwenden?

(b) Für Personen mit „Extremwerten“, d.h. für Personen, die alle Items oder kein einziges Item gelöst haben bzw. alle Items mit den stärksten oder den schwächsten Antwortalternativen beantwortet haben, sind keine Personenparameter ermittelbar.

(c) Die Gültigkeit des Raschmodells für einen gegebenen Datensatz wird widerlegt, wenn der Stichprobenumfang sehr groß wird.

(d) Das Raschmodell ist für zwei-kategoriale Antwortformate brauchbar, nicht aber für mehr-kategoriale Antwortformate

(e) Bei Verwendung des Raschmodells erscheint es unzulässig, wenn anstelle der Rasch-Personenparameter (s.o.) die ansonsten üblichen Summenscores verwendet werden.

(f) Für die Berechnung der Raschskalierung gibt es nur schwierig anwendbare EDV-Programme

Kurze Demonstration von Winmira-Ergebnissen

Lebenszufriedenheit

Deprimiertheit

Neben den theoretischen Vorteilen hat die Rasch-Skalierung auch praktische Vorteile

(a) Neben dem Gesamtfit für eine Skala
und den Fit-Indices für jedes Item (die sich anstelle der Trennschärfe für Itemselektionen verwenden lassen),
(b) Gibt es auch Fit-Indices für die Personen, was ermöglicht sog. „Ausreißer“ zu ermitteln und auf fehlerhafte Dateneingaben zu prüfen.
(c) Es ist kein Problem, unterschiedlich lange Antwortkategorie-Vorgaben für die Items einer Skala zu verwenden („partial credit model“).
(d) Es gibt Hinweise darauf, ob die verwendeten Antwortkategorien angemessen sind, bzw. ob man lieber bestimmte Antwortkategorien zusammenfassen sollte (dies wird pro Item untersucht).
(e) Wegen der meist sehr hohen Korrelation zwischen den Rasch-Personen-Parametern und den Summenscores können der Einfachheit halber die Summenscores verwendet werden, denn sie sind (wenn die Items Rasch-geeignet sind) erschöpfende Statistiken
(f) Mit der Raschskalierung geprüfte Skalen lassen sich für ein „computeradaptives Testen“ verwenden

Empfehlung:
Bei Skalen-Überprüfungen
klassischer Art
die Raschskalierung
Wenigstens „mitführen“

Beispiel:

Furthermore we conducted Rasch analyses of the PGSI subscale, total scale and items using the program WINSTEPS. The Rasch reliability coefficients are satisfactory.

Regarding the Rasch model fit criteria, our PGSI values are quite near to the optimal value 1.00 and between 0.4 and 1.2, it can be concluded that our PGSI is acceptable also on the base of Rasch model criteria (see Wright, Linacre, Gustafson, & Martin-Loff, 1994).

Additionally we tested according to Wright & Masters (1982) the appropriateness of the seven-category rating scale for the PGSI items; for only eight of the 36 items the WINSTEP program suggested in each case collapsing two of the seven categories. In order to preserve uniqueness for the whole PGSI we decided to keep the use of seven categories unchanged.

Zusätzlich ließen sich noch die Korrelationen der Rasch-Personenparameter mit den („klassischen“) Summenwerten berichten, die Korrelationen liegen meist über r=0.95, womit sich begründen lässt, dass die Verwendung der Summenwerte in den nachfolgenden statistischen Auswertungen gerechtfertigt erscheint (wenn die Skalen den Raschmodell-Prüfkriterien genügen).

Ende