Ein SAS-Modul zur Konstruktion psychologischer Tests nach dem Rasch-Modell

Stefan Klein

Humboldt-Universität zu Berlin, Institut für Psychologie
Telefon: 030 / 285165226
eMail: stefan.klein@rz.hu-berlin.de

Abstract

Dieser Text stellt ein SAS-Macro zur Konstruktion psychologischer Tests vor. Dazu wird zunächst anhand eines Beispiels demonstriert, wie ein psychologischer Test aussieht, und nach welchen Prinzipien ein solcher Test konstruiert wird. Es werden dazu die Gütekriterien der Testkonstruktion und ihre Bedeutung für die Aufgabenselektion dargestellt. Weiterhin werden die Ansätze der klassischen und der probabilistischen Testtheorie kurz dargestellt. Schließlich wird der Funktionsumfang des SAS-Macros beschrieben, sowie kurz auf geplante Erweiterungen eingegangen.

1. Einleitung

In diesem Beitrag wird der Inhalt des Macro-Pakets „Analysis of Change“ vorgestellt. Dies ist ein Macro-Paket, das zur Aufgabenselektion bei der Konstruktion psychologischer Tests verwendet werden kann.
Grundidee der Aufgabenselektion ist, daß ein Psychologe aus verschiedenen ihm vorliegenden Items diejenigen auswählt, die am besten zur Messung einer bestimmten Fähigkeit geeignet sind.
Im Macro „ANALYSIS of Change“ werden dazu sowohl Aspekte der klassischen Testtheorie (=KTT) als auch Aspekte der probabilistischen Testtheorie (=PTT) verwendet. Außerdem kann das Macro-Paket die Parameter für gewisse Modelle der Veränderungsmessung schätzen (zu diesen Modellen: vgl. Rost [1996]). Neu an diesem Macro-Paket ist vor allem die Bereitstellung einer SAS-Oberfläche für die beiden Ansätze der psychologischen Testtheorie. Ziel war es dabei, eine Oberfläche mit relativ einfacher Benutzerführung zu entwickeln, die es erleichtert, die zur Aufgabenauswahl notwendigen Schritte durchzuführen. Weiterhin soll das Programm es einem größeren psychologischen Benutzerkreis ermöglichen, komplizierte Item-Response-Modelle für die Veränderungsmessung zu verwenden. Das Neue an diesem Programm ist dabei – auch im Vergleich mit nicht-SAS-Programmen – v.a. die gemeinsame Verwendung von klassischer Testtheorie und Item-Response-Modellen in einem Programm.
Im Sommersemester 2000 wird dieses Macro voraussichtlich im Rahmen eines Seminars zur Modellbildung in der Psychologie als Lernsoftware für die Anwendung von Rasch-Modellen verwendet.
Weiterhin sollen die Ideen der Aufgabenselektion an einem kleinen Beispiel von 10 Items demonstriert werden. Die Items, an denen dies geschieht, sind Teil einer Onlinebefragung zur Untersuchung der Internetkenntnisse bei Schülern der Jahrgangsstufen 7 bis 13, an der 758 Schüler teilnahmen. 422 Schüler beantworteten die folgenden 10 Items vollständig:

1. Kannst Du zwischen verschiedenen WWW-Browsern wählen ?
2. Macht Dir die Benutzung des WWW-Browsers Spaß ?
3. Kennst Du alle Funktionen deines Browsers ?
4. Nutzt Du die Möglichkeit, eigene Einstellungen am Browser vorzunehmen ?
5. Hast Du eine eigene Homepage ?
6. Hast Du eine eigene Favoriten- oder Benchmarkliste ?

7. Kannst Du zwischen verschiedenen E-Mail-Programmen wählen ?
8. Macht Dir das Versenden von E-Mails Spaß ?
9. Kennst Du alle Funktionen Deines E-Mail-Programms ?
10. Hast Du ein eigenes Adressbuch auf dem Computer ?

Diese 10 Fragen werden als Datengrundlage für das Demonstrationsbeispiel benutzt.

Das Macro-Paket verwendet intern vor allem PROC IML zur Berechnung der Schätzungen für das Rasch-Modell. Die Oberfläche und große Teile der Berechnung der Koeffizienten der klassischen Testtheorie wurden mit Hilfe von SAS-Macros und DATA-Step-Programmierung implementiert.
Die Dateneingabe erfolgt –wenn die Daten im Excel 5.0-Format vorliegen– über PROC ACCESS, bzw. – wenn die Daten schon SAS-Format besitzen – über den DATA-Step. Weitere Eingabemöglichkeiten werden in Zukunft ebenfalls implementiert werden, wie z.B. die Dateneingabe aus SPSS.
Das Macro-Paket besteht aus den beiden SAS-Programmen „anacha10“ und „setup_an“. Das letztere Programm definiert eine Macrobibliothek und eine IML-Bibliothek im Verzeichnis SASUSER, die die benötigten Macros und IML-Prozeduren enthalten. Das Programm „anacha10“ ruft diese Prozeduren und Macros auf, und ist für die Datenausgabe zuständig.
Zur Istallation des Macro-Pakets genügt es, die Dateien auf die Festplatte zu kopieren, und dann „setup_an“ zu starten. Zur Verwendung des Macro-Pakets reicht das Programm „anacha10“ aus.

2. Was sind psychologische Tests ?

a) Einführung

Ein wichtiges Ziel der Psychologie liegt in der Messung der Stärke von Persönlichkeitsausprägungen. Beispiele für solche Persönlichkeitsausprägungen sind Konstrukte wie Intelligenz, Aggressivität etc.
Psychologische Tests werden zu verschiedensten Zwecken eingesetzt, wie z.B.

- die Messung der Intelligenz (IQ-Tests)
- Leistungstests zur Aufnahme an Universitäten
- Tests zur Messung der Kundenzufriedenheit
- Messung der Lebensqualität bei Krebspatienten.

Als Meßinstrument werden in der Regel Fragebögen verwendet, die aus mehreren Items (oder: Aufgaben) bestehen. Diese Items können verschiedenartig aufgebaut sein (vgl. auch z.B. Lienert/Raatz [1994]):
Oft verwendet man als Items Aufgaben, die ein Proband lösen muß. Inhalt und Art dieser Aufgaben hängt dann von der Art der Fähigkeit ab, die gemessen werden soll. Beispielhaft dafür sind Rechenaufgaben, Wortergänzungsaufgaben, usw. (vgl. Lienert/Raatz [1994]).
Oft besteht ein Item jedoch lediglich aus einer Frage, zu der der Proband eine Antwort aus zwei oder mehr vorgegebenen Alternativen auswählen muß.

b) Dichotome und Polytome Items

Diese Begriffe werden anhand eines Beispiels erläutert:
Für einen Test zur Ermittlung der mathematischen Begabung einer Person werden den Probanden mehrere Rechenaufgaben dargeboten. Falls man hierbei nur unterscheidet, ob diese Aufgaben gelöst wurden oder nicht, spricht man von dichotomen Items. Wenn man dagegen auch verschiedene Teillösungen berücksichtigt, spricht man von polytomen Items.

Dies kann am Beispiel der (Rechen)aufgabe „Berechne $\sqrt{3*5}$ “ näher erläutert werden: Man kann hier drei Fähigkeitsstufen der Rechenfähigkeit unterscheiden (die hier mit 0,1,2 kodiert werden):

- Personen die die Aufgabe vollständig lösen, wird der Ausprägungsgrad 2 zugewiesen
- Personen, die zwar 3*5 richtig berechnen können, die aber keine Wurzel ziehen können, wird der Fähigkeitsgrad 1 zugeordnet
- Personen, die weder die Wurzel ziehen können, noch die Multiplikation 3*5 berechnen können, ordnet man den Ausprägungsgrad 0 zu.

Ein solches polytomes Item kann als ordinal skalierte Größe aufgefaßt werden, bei der zwischen den verschiedenen Ausprägungsgraden eine Ordnungsrelation besteht. Meistens kodiert man polytome Items mit den Zahlen 0, 1, ... Dabei steht 0 immer für den geringsten Ausprägungsgrad einer Fähigkeit.

c) Trennschärfe und Schwierigkeit eines Items

Hierzu verwenden wir wieder das Beispiel eines Tests für mathematische Begabung, der – wie vorher dargestellt – aus einer Reihe von Rechenaufgaben besteht.
Rechenaufgaben kann man leicht eine Aufgabenschwierigkeit zugeordnen, die ausdrückt, wie häufig eine Aufgabe gelöst wird. Meistens wird die Schwierigkeit eines Items daher als Anteil der Probanden in einer Stichprobe operationalisiert, die ein bestimmtes Item gelöst haben.
Von der Schwierigkeit einer Aufgabe zu unterscheiden ist die Trennschärfe. Dieser Koeffizient drückt aus, wie gut ein Item Probanden mit hoher Fähigkeitsausprägung von Probanden mit niedriger Fähigkeitsausprägung unterscheiden kann. Ein Item mit hoher Trennschärfe wird nur von den „guten“ Probanden gelöst , nicht jedoch von den „schlechten“.
Oft wird die geringe Trennschärfe jedoch v.a. durch ganz andere Faktoren beeiflußt. So würde z.B. die Verwendung eines verbalen Items in obigem Mathematik-Test zu einer relativ geringen Trennschärfe führen, da zur Lösung eines verbalen Items auch das Sprachverständnis gefragt ist, und das Item daher zwei Fähigkeiten gleichzeitig messen würde.
Trennschärfe und Schwierigkeit eines Items kann man sich anhand des Itemtypus „Aufgabe“ gut veranschaulichen. Oft bestehen Items jedoch nicht aus solchen Aufgaben, sondern aus Fragen, auf die mit mehreren vorgegebenen Kategorien geantwortet werden kann. Bei Items dieser Art werden die Begriffe Schwierigkeit und Trennschärfe zwar ebenfalls verwendet, sind aber nicht mehr so anschaulich zu interpretieren.

d) Reliabilität und Validität

Wie oben erwähnt, ist ein Test eine Zusammenstellung mehrerer Items. Ziel ist es, einen Meßwert zu erhalten, der etwas über die zu untersuchende Fähigkeit aussagt. Dazu verwendet man zumeist die ungewichtete Summe über alle Itemantworten eines Probanden, den Summenscore. Um diesen Summenscore bilden zu können, muß eine Codierung der Antworten vorausgesetzt werden, bei der eine Summation überhaupt möglich ist. Meist verwendet man für die einzelnen Items dichotome oder ordinale Antwortformate.
Um den Summenscore als Fähigkeitsmeßwert verwenden zu können, sind jedoch bestimmte Voraussetzungen bezüglich der Items zu erfüllen. Dazu gibt es (im wesentlichen) zwei Ansätze: die klassische und die probabilistische Testtheorie.
Allgemein unterscheidet man in diesem Rahmen die Begriffe der Validitätt und der Reliabilität.
Reliabilität drückt die Zuverlässigkeit der Messung durch einen psychologischen Test aus und kann mit dem Begriff Meßgenauigkeit gleich gesetzt werden. Reliabilität eines Items wird oft über die Korrelation zwischen einem einzelnen Item und dem Gesamtmeßwert operationalisiert. Je größer diese Korrelation ist, desto größer ist auch die Reliabilität der Aufgabe.

Daher wird der Begriff „Reliabilität einer Aufgabe“ auch oft als Synonym für die Trennschärfe einer Aufgabe verwendet.
Die Reliabilität eines Tests läßt sich am besten anhand der Retest-Reliabilität veranschaulichen: Ein Test ist dann retest reliabel, wenn er (falls er dem gleichen Probanden nach angemessener Zeitspanne wieder vorgelegt wird) zum gleichen Testergebnis führt.
Validitätskoeffizienten untersuchen hingegen, ob ein Test wirklich die interessierende Fähigkeit mißt, oder etwas völlig anderes.
Um Validität zu bestimmen gibt es zwei Hauptansätze:
Zum einen kann man die Validität eines Items (oder eines gesamten Tests) durch den Vergleich mit einem Außenkriterium bestimmen (=externe Validität). Ziel wäre es in diesem Fall, daß das Item (bzw. der gesamte Test) hoch mit dem Außenkriterium korreliert. Im Beispiel eines Tests auf mathematische Begabung könnte dieses Außenkriterium z.B. die Abiturnote in Mathematik sein.
Andererseits kann man aber auch die „interne Validität“ bestimmen. Diese leitet sich aus bestimmten Beziehungen her, die zwischen den einzelnen Items gelten sollten, wenn der Test das mißt, was er messen soll. So sollten z.B. Personen, die schwierige Items lösen können, auch in der Lage sein, die leichteren Items zu lösen.
Neben diesen Hauptansätzen gibt es noch mehrere andere Definitionen des Begriffs „Validität“. Heute werden zur Absicherung der Validität eines Tests oft Strukturgleichungsmodelle verwendet.
Bei der Itemselektion im Rahmen eines Tests sucht man Items, die eine möglichst hohe Reliabilität bei gleichzeitig hoher Validität besitzen (vgl. auch Cronbach [1990]).
Zur Messung der Reliabilität eines Items verwendet man meist die Korrelation mit dem Summenscore. Diese Operationalisierung bewirkt, daß der Test nur Items enthält, die sich ähnlich wie das Ergebnis des Gesamttests verhalten und somit eine hohe Trennschärfe besitzen.
Validität wird hingegen meist als Korrelation zu einem Außenkriterium operationalisiert.

3. Die klassische Testtheorie

a) Das Modell der klassischen Testtheorie

Das Modell der klassischen Testtheorie leitet sich aus der Faktorenanalyse ab. Dabei wird meist davon ausgegangen, daß ein Test genau ein Merkmal messen sollte, wobei folgende Grundgleichung zur Erklärung des gemessenen Testwertes zugrundegelegt wird:

$$X = T + E$$

Hierbei bedeutet
E: Fehlervariable
T: Wahrer Wert der Fähigkeit
X: gemessener Wert eines Items
Dies ist ein Ein-Faktoren-Modell mit dem (allen Variablen gemeinsamen) Faktor T. Die Grundgleichung wird für jedes Item vorausgesetzt. Damit ergibt sich eine weitere Möglichkeit zur Validierung eines Tests: Durch eine Faktorenanalyse kann man überprüfen, ob ein Test wirklich nur eine Fähigkeit (=Trait) mißt. Falls dies nicht der Fall ist, ist der Test inhomogen und daher –falls die zu untersuchende Eigenschaft hinreichend genau definiert werden kann– nicht valide für die Messung der interessierenden Eigenschaft.
Erweiterungen dieses Konzepts führen zu latenten Strukturgleichungsmodellen. Hier wären z.B. die Konzepte der Tau-Kongenerizität und der Tau-Äquivalenz zu nennen (vgl. auch Steyer/Eid [], Rost [1999]).
Als Kritik an diesem klassischen Konzept wird oft angeführt, daß die Modellgleichung $X = T + E$ auf dichotome Items nur schwer anwendbar ist. Diese Kritik führt u.a. zur Entwick-

lung der sogenannten Item-Response-Modelle, deren bekanntestes das Rasch-Modell ist (vgl. Rost [1999]).

b) Konzepte der klassischen Testtheorie im vorgestellten Macro

Da die klassische Testtheorie immer noch Standard in der psychologischen Forschung ist, enthält das Macro-Paket auch Prozeduren zur Berechnung von Reliabilitäts- und Validitätsmaßen, die zur Itemselektion nach klassische Methode verwendet werden. „Analysis of Change“ berechnet auf Macro-Basis verschiedene – dem Skalenniveau angepaßte – Korrelationskoeffizienten. Zur Bestimmung der Reliabilitäten werden

- bei dichotomen Items punktbiseriale Korrelationen mit einem um den Wert des Items korrigierten Summenscore berechnet, sowie
- bei polytomen Items Kendalls τ_b.

Weiterhin werden die Korrelationen zwischen den Items ausgegeben. Dazu wird im dichotomen Fall Yules' Q-Koeffizient berechnet, im polytomen Fall wiederum Kendalls τ_b.

Zur Messung der Validität wird Kendalls τ_b zwischen den Items und einem Außenkriterium verwendet.

Weiterhin wird– als Anwendung von PROC CORR und zur Vervollständigung des Angebots an Korrelationskoeffizienten – Cronbachs Alpha-Maß als Maß für die Reliabilität des Gesamttests ausgegeben.

Die eigentliche Itemselektion muß mit Hilfe der hier berechneten Itemkennwerte vorgenommen werden. Dazu kann man entweder per Augenschein Items mit zu geringer Reliabilität / Validität aussortieren oder eine schematische Selektionstechnik verwenden. Die gebräuchlichste dieser Selektionstechniken –die sog. Gulliksen-Technik– verwirft Items mit einem niedrigen Verhältnis

$$\frac{\text{Validität}}{\text{Reliabilität}}.$$

Außerdem werden solche Items nicht berücksichtigt, bei denen Reliabilität und/oder Validität kleiner als 0 sind. (Zur Erinnerung: die verwendeten Korrelationskoeffizienten haben einen Wertebereich zwischen –1 und +1). Zur Anwendung der Gulliksen-Technik im Macro muß die Zahl der Items angegeben werden, die der Test enthalten soll; das Macro sortiert dann automatisch die besten Items nach Gulliksen aus.

Im Bereich der klassischen Testtheorie gibt das Modul einen Datensatz für die Reliabilitäts- und Validitätskoeffizienten, sowie für die Inter-Item-Korreltionen aus, Weiterhin wird eine Tabelle für die Ergebnisse der Gulliksen-Selektion ausgegeben.

c) Demonstration am Beispieldatensatz

Mit „Analysis of Change“ konnten Reliabilitätskoeffizienten zwischen 0.06 bei Aufgabe 8 und 0.39 bei Aufgabe 3 berechnet werden. Dies sind relativ geringe Reliabitäten, was aber unter Umständen dadurch zu erklären ist, daß dichotome Items meist eine relativ geringe Trennschärfe besitzen. Im Ernstfall der Konstruktion eines Tests würde man hier zumindest das am wenigsten trennscharfe Item aussortieren, zumal dieses Item auch eine relativ geringe Validität besitzt.

Die Validitäten schwankten zwischen 0.09 bei Aufgabe 8 und 0.27 bei Aufgabe 3. Diese Validitäten wurden mit Hilfe eines 7-Stufigen Außenkriteriums berechnet, nämlich dem in der gleichen Befragung gestellten Item

- Wie oft bist Du durchschnittlich im Internet ?

Die Antworten auf der 7-stufigen Skala des Außenkriteriums umfaßten den Bereich von „mehrmals täglich“ bis zu „seltener als einmal im Monat“.
Auch die Validität unserer Beispieldaten ist als recht gering zu betrachten.
Nach der Gulliksen-Selektion wurden – um eine Höchstzahl von 8 Items im Test zu erhalten – die Items 4 und 7 entfernt, da das obengenannte Verhältnis bei diesen Items am niedrigsten war.
Allerdings ist das Verhältnis von Validität zu Reliabilität bei allen Items des Beispiels ziemlich ähnlich. Bezüglich der verwendeten Gütekriterien sind die Items daher recht homogen auf eher niedrigem Niveau.
Die Schwierigkeiten der Items reichen von 0.26 bei Item 5 bis zu 0.86 bei Item 2 und Item 8. Da man meistens versucht, mit einem Test einen großen Schwierigkeitsbereich abzudecken, würde man diese – bezüglich der Schwierigkeit – extremen Items als Bestandteil des Tests behalten.

4. Die probabilistische Testtheorie

a) Der theoretische Ansatz des Rasch-Modells

Als Modell der probabilistischen Testtheorie wird hier einzig das sog. Rasch-Modell (Rasch 1960) behandelt. Allen Modellen der probabilistischen Thesttheorie gemeinsam ist ihre Verwandtschaft zu den Latent-Trait-Modellen der Kontingenztafelanalyse (vgl. Andersen 1989). Diese Modelle versuchen die Abhängigkeiten zwischen verschiedenen (kategorial oder ordinal skalierten) manifesten Variablen durch eine latente Variable zu erklären, die den beobachteten Variablen zugrundeliegt. Im Zusammenhang mit der psychologischen Testtheorie wird meist angenommen, daß diese latente Variable metrisch skaliert ist; es gibt jedoch auch Modelle, die kategorial skalierte latente Variable verwenden (vgl. dazu auch Rost 1996; Rost 1999).
Das Rasch-Modell ist das wichtigste Modell der probabilistischen Testtheorie. Es existieren Varianten dieses Modells sowohl für polytome als auch für dichotome Items. Hier wird ausschließlich die dichotome Variante beschrieben (vgl. dazu auch Rost [1996]; Rost [1999]).
Das Rasch–Modell versucht, die Wahrscheinlichkeit zu modellieren, mit der eine Person S_v ein dichotomes Item I_j löst. Die Lösungswahrscheinlichkeit hängt nun ab von der Fähigkeit λ_v der Person und der Schwierigkeit β_j des Items ab. Die Lösungswahrscheinlichkeit besitzt folgende Form:

$$P(X_{vj} = x_{vj}) = \frac{\exp[x_{vj}(\lambda_v - \beta_j)]}{1 + \exp(\lambda_v - \beta_j)}$$

Die Zufallsgröße X_{vj} modelliert dabei das Antwortverhalten der Person S_v bei Frage I_j. Wie immer steht x_{vj} dabei für die tatsächliche Antwort von Person v bei Item i.
Die Werte für die Parameter λ_v und β_j werden aus den Daten mittels (Conditional-) Maximum-Likelihood-Schätzung bestimmt. Neben dem Summenscore wird auch λ_v als Schätzwert für die Fähigkeit einer Person verwendet.
Das Rasch-Modell wurde konstruiert, um sicherzustellen, daß ein psychologischer Test folgende wünschenswerte Eigenschaften besitzt:

- Spezifische Objektivität: Das Meßergebnis für eine Person soll nicht davon abhängen, welches Meßinstrument zur Messung der Fähigkeiten verwendet wird. Dies bedeutet: Zwei spezifisch objektive Tests mit unterschiedlichen Items (die aber die gleiche Fähigkeit messen) kommen zu gleiche Schätzwerten für die Fähigkeit.

- Suffizienz des Summenscores: Der Summenscore enthält die ganze Information über die Fähigkeit einer Person, die in dem Test vorhanden ist. Anschaulich bedeutet dies, daß alle Antwortmuster mit dem gleichen Summenscore dieselbe Information über die interessierende Fähigkeit besitzen. Nur in diesem Fall ist es überhaupt sinnvoll, den Summenscore als Maß für die Fähigkeit einer Person zu verwenden.
- Lokale stochastische Unabhängigkeit: Unter der Bedingung konstanter Fähigkeitswerte ist das Ereignis „Lösen des Items i“ unabhängig vom Ereignis „Lösen des Items j“:

  $$P(X_{vi} = 1 | \lambda_v) = P(X_{vj} = 1 | \lambda_v)$$

Die letztere Eigenschaft impliziert, daß es keine verborgenen Abhängigkeiten zwischen den Items gibt, wenn man konstante Fähigkeitswerte voraussetzt. Lokale stochastische Unabhängigkeit ist eine Eigenschaft, die für jeden psychologischen Test wünschenswert ist.
Man kann zeigen, daß psychologische Tests, die diese Eigenschaften besitzen, dem Rasch-Modell genügen (vgl. Fischer, [1995a]). Somit liegt hier ein überprüfbares Kriterium vor, wann aus einer gegebenen Menge von Items ein psychologischer Test mit sinnvollen Eigenschaften konstruiert werden kann.
Neben diesem (klassischen) Rasch-Modell gibt es verschiedene Erweiterungen. Eine dieser Erweiterungen, dient (siehe auch Fischer [1995b]) als Modell für die Messung von Veränderungen. Vorteil dabei ist, daß durch die Verwendung latenter Größen der wahre Wert der Veränderung mit größerer Sicherheit gemessen werden kann. Diese Erweiterung zur Veränderungsmessung ist ebenfalls Teil des Macro-Pakets „Analysis of Change“.
In der Praxis wurde das Rasch-Modell bisher hauptsächlich als „Hilfsinstrument“ zur Gewinnung homogenerer Tests eingesetzt, während der Hauptteil der Testkonstruktion nach dem klassischen Konzept erfolgte (vgl. auch Lienert, [1994]). Dies ist einerseits auf den recht komplizierten mathematischen Hintergrund zurückzuführen, andererseits auf die Tatsache, daß es oft schwierig ist, genügend Items zu finden, die die „schwereren“ Voraussetzungen einer Rasch-Skala erfüllen (vgl. auch Rost,[1999]).

b) Die Implementierung des Rasch-Modells im Macro-Paket „Analysis of Change“

Wie der Name des Macro-Pakets ausdrückt, ist die Macro-Lösung für die Schätzung von IRT-Modellen in der Veränderungsmessung konzipiert. Allerdings ist damit auch die Schätzung von Modellen für nur einen Zeitpunkt möglich.

Das Macro-Paket kann die Parameter zweier unterschiedlicher Modellansätze schätzen:

Im Modellansatz „RSM“ (für „Rating Scale Model“) wird (bei mindestens zwei Zeitpunkten) eine Veränderung geschätzt, die für alle Items (und alle Personen) gleich ist. Falls mehrere Subpopulationen vorhanden sind, wird für jede Subpopulation ein eigener Veränderungsparameter geschätzt.
Dieser Ansatz ist ein Spezialfall des sog. „Linear Rating Scale-Modells“ (vgl. Fischer/Ponocny [1991]).
Falls nur ein Zeitpunkt verwendet wird, geht dieses Veränderungsmodell in das normale Rasch-Modell über.
Das zugrundegelegte Modell für ordinale Items ist das sogenannte Rating-Scale-Modell (vgl. Rost [1996]). Im Fall von dichotomen Antworten erhält man als Spezialfall dieses Rating-Scale-Modells das normale Rasch-Modell für dichotome Items.

Im Modellansatz „LTM“ (= Logistic Trend Model, Bezeichnung wie im Macro-Paket) wird bei mindestens zwei Zeitpunkten eine Veränderung geschätzt, die für jedes Item einen anderen Wert annehmen kann.
Bei Verwendung für nur einen Zeitpunkt entspricht dieses Modell dem oben beschriebenen ersten Modellansatz. Anzumerken wäre noch, daß mit dieser zweite Modellansatz auch für das „Linear Logistic Model with Relaxed Assumptions“ (= LLRA, Fischer [1995b]) verwendet werden kann.
Als Schätzverfahren für die Veränderungs- und Fähigkeitsparameter wird bei beiden Veränderungsansätzen die Conditional-Maximum-Likelihood-Schätzung verwendet. Die Schätzung der Fähigkeitsparameter basiert auf (normaler) Maximum-Likelihood-Schätzung.
Das Macro-Paket gibt eine Tabelle für die Schätzung der Fähigkeitsparameter und eine Tabelle für die Schätzung der Itemschwierigkeiten aus.
Als Test für die Modellanpassung ist bisher nur der Andersen-Test implementiert. Dieser Test teilt die Stichprobe nach den Testergebnissen in mehrere (Score-) Gruppen auf und vergleicht die Parameterschätzungen in diesen Gruppen mit den Parameterschätzungen im gesamten Modell.

c) Beispieldemonstration

Für den obigen Datensatz wurden die Parameter für ein Rasch-Modell mit einem Zeitpunkt und 10 Items geschätzt.
Die Itemschwierigkeiten schwankten zwischen –1,58 bei Item 5 und 2.01 bei Item 2. Interessant hierbei ist, daß die schwierigsten Items aus der klassischen Testtheorie (nämlich die Items 2 und 8) auch nach dem Rasch-Modell am schwierigsten sind. Item 5 wiederum war auch in der KTT das leichteste Item.
Die Itemselektion erfolgt im Rasch-Modell nicht mehr über Reliabilitäten und Validitäten: Statt dessen werden verschiedene Itemkombinationen ausprobiert, bis das Rasch-Modell nicht mehr durch den Modellgeltungstest abgelehnt wird. Normalerweise wird man dabei Items bevorzugen, die eine relativ ähnliche (klassisch gemessene) Trennschärfe besitzen. Im Gegensatz zur klassischen Testtheorie wird man über das Rasch-Modell aber stets tests mit relativ homogenen Trennschärfen erhalten. In unserem Fall konnte das Rasch-Modell nicht verworfen werden; man könnte also alle Items in eine (gemeinsame) Rasch-Skala aufnehmen. Dieses Ergebnis korrespondiert mit den Ergebnissen der klassischen Testtheorie: auch bei der klassischen Itemanalyse stellte sich, der Beispielsdatensatz als recht homogen bezüglich Trennschärfe und Validität heraus.
Interessant ist weiterhin die Tabelle der geschätzten Fähigkeitswerte nach dem Rasch-Modell.

Summenscore	**Zugeordnete Schwierigkeit**
Score 1	-2.67
Score 2	-1.64
Score 3	-0.92
Score 4	-0.32
Score 5	0.19
Score 6	0.70
Score 7	1.23
Score 8	1.85
Score 9	2.74

Tab. 1

Für die Scores 0 und 10 kann –aus prinzipiellen Gründen– im Rasch-Modell keine Fähigkeit gemessen werden. Vorteil dieser Art von Fähigkeitsschätzung gegenüber dem Summenscore ist, daß die geschätzten Fähigkeiten eine Intervallskala bilden.

5. Diskussion

Das vorgestellte Macro-Paket ist Teil der Dissertation des Referenten. Es umfaßt im gegenwärtigen Zeitpunkt die Fähigkeit, die Parameter für bestimmte Item-Response-Modelle zur Veränderungsmessung zu schätzen, sowie die Modellgeltung mit dem Andersen-Test zu überprüfen. Ziel ist ein Programm, mit dem die wichtigsten Aspekte sowohl der klassischen als auch der probabilistischen Testtheorie abgedeckt werden.
Allerdings ist dies nur eine erste Version des Macro-Pakets. Weitere Features sollen im Zuge der Dissertation des Dozenten mitaufgenommen werden. Dazu zählen:

- graphische Tools zur Analyse der Residuen in Item-Response-Modellen
- Modellgeltungstests für einzelne Items
- Bootstrap-Tests für kleine Stichproben
- Reliabilitätsschätzung mit dem Rasch-Modell
- Kriterien für die Auswahl von Items im Rasch-Modell
- erweiterte Eingabemöglichkeiten

Für alle interessierten Psychologen soll das Macro-Paket darüberhinaus im Internet zur Verfügung gestellt werden. Informationen dazu können ab Ende April 2000 auf der Homepage des Referenten (http://www.psychologie.hu-berlin.de/met/met15.htm) abgerufen werden.

Literaturverzeichnis

Andersen, E.B. (1989). The Statistical Analysis of categorical Data. Berlin: Springer.

Cronbach, Lee J. (1990). Essentials of Psychological Testing, 5th Ed. New York: HarperCollins.

Fischer, Gerhard H. (1995a). Derivations of the Rasch Model. aus: Fischer, G.H. / Molenaar, I.W. Rasch Models Foundations, Recent Developments, and Applications. S.15-39. Berlin: Springer.

Fischer, Gerhard H. (1995b). Linear logistic models for change. aus: Fischer, G.H. / Molenaar, I.W. Rasch Models Foundations, Recent Developments, and Applications. S.157-180. Berlin: Springer.

Fischer, Gerhard H. / Ponocny, I. An Extension of the partial credit model with application to the measurement of change. Psychometrika Vol. 59, S.177-192.

Lienert, G.A. / Raatz, U. (1994). Testaufbau und Testanalyse. 5. Auflage. Weinheim: Beltz.

Rasch, G. (1960). Probabilistic Models for some intelligence and attainment tests. Copenhagen: Nielsen & Lydiche.

Rost, Jürgen (1996). Lehrbuch Testtheorie, Testkonstruktion. Bern: Huber.

Rost, Jürgen (1999). Was ist aus dem Rasch-Modell geworden?. Psychologische Rundschau, Vol. 50, Band 3, S.140-153

SAS Institute (1990). SAS/IML Software: Usage and Reference Version 6. 1st Ed. Cary, NC

Steyer, R. / Eid, M. (1993). Messen und Testen. Berlin: Springer.