Koller / Alexandrowicz
Hatzinger

Das Rasch Modell in der Praxis

Eine Einführung in eRm

facultas wuv UTB

Eine Arbeitsgemeinschaft der Verlage

Böhlau Verlag · Wien · Köln · Weimar
Verlag Barbara Budrich · Opladen · Farmington Hills
facultas.wuv · Wien
Wilhelm Fink · München
A. Francke Verlag · Tübingen und Basel
Haupt Verlag · Bern · Stuttgart · Wien
Julius Klinkhardt Verlagsbuchhandlung · Bad Heilbrunn
Mohr Siebeck · Tübingen
Nomos Verlagsgesellschaft · Baden-Baden
Orell Füssli Verlag · Zürich
Ernst Reinhardt Verlag · München · Basel
Ferdinand Schöningh · Paderborn · München · Wien · Zürich
Eugen Ulmer Verlag · Stuttgart
UVK Verlagsgesellschaft · Konstanz, mit UVK / Lucius · München
Vandenhoeck & Ruprecht · Göttingen · Oakville
vdf Hochschulverlag AG an der ETH Zürich

Ingrid Koller, Rainer Alexandrowicz,
Reinhold Hatzinger

Das Rasch-Modell in der Praxis

Eine Einführung mit eRm

facultas.wuv

Bibliografische Information Der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie;
detaillierte bibliografische Daten sind im Internet über
http://dnb.d-nb.de abrufbar.

© 2012 Facultas Verlags- und Buchhandels AG
facultas.wuv, Wien
Alle Rechte vorbehalten.

Druck und Bindung: CPI-Ebner & Spiegel, Ulm

UTB-Band-Nr.: 3786
ISBN 978-3-8252-3786-8

Für Reinhold

Vorwort

Bei vielen wissenschaftlichen Untersuchungen geht es darum herauszufinden, wie sehr Personen über bestimmte Eigenschaften oder Fähigkeiten verfügen oder wie stark ausgeprägt bestimmte Einstellungen oder Neigungen sind. Zur Beantwortung solcher Fragen werden als Messinstrumente oft Fragebögen oder Tests verwendet. Ein wichtiger Aspekt hierbei ist die Qualität der Messinstrumente. Sind die Fragen eines Fragebogens oder die Aufgaben eines Tests nicht gut konstruiert, so werden auch die Ergebnisse mangelhaft sein. Man benötigt also Methoden, um die Qualität solcher Instrumente evaluieren und sicherstellen zu können.

Eine der wichtigsten dieser Methoden besteht in der Anwendung des Rasch Modells. Das Rasch Modell eignet sich nicht nur zur Analyse oder Konstruktion von Fragebögen und Tests an sich, sondern bewährt sich auch in zahlreichen anderen Anwendungsgebieten. Dies reicht von sehr praktischen Fragen nach Unterschieden zwischen Personengruppen (z. B. Zielgruppensegmentierung im strategischen Marketing) oder der Prognose von zukünftigem Verhalten (z. B. in der Psychotherapie) bis zur Evaluierung theoretischer Hintergrundmodelle (z. B. in der medizinischen Diagnostik).

Dieses Buch konzentriert sich auf die Anwendung des dichotomen Rasch Modells. (Dichotom bedeutet, dass die Antworten auf Fragen, die Lösung von Aufgaben eines Tests oder das Auftreten eines bestimmten Verhaltens in zwei Kategorien erfolgt, d. h. „ja/nein“, „stimme zu/nicht zu“, „richtig/falsch“ oder „tritt auf/tritt nicht auf“.) Wir haben uns zum Ziel gesetzt, die theoretischen Grundlagen verständlich darzustellen sowie die Anwendung des Rasch Modells einerseits für Studierende zu erleichtern und andererseits Praktikern, die das Rasch Modell für eigene Studien anwenden möchten, ein nachvollziehbares Konzept zu liefern.

Das Buch gliedert sich im wesentlichen in drei Teile. Nach einer Einführung in die wichtigsten Qualitätsmerkmale von Messungen in den ersten beiden Kapiteln werden die statistischen Grundlagen des Rasch Modells und der Überprüfung der Qualitätskriterien in den Kapiteln 3 bis 5 dargestellt. Während die wesentlichen Standard-Überprüfungsmöglichkeiten in Kapitel 4 behandelt werden, gibt Kapitel 5 eine Darstellung neuester Methoden der

Qualitätskontrolle des Modells, die sich auch für kleinere Stichproben eignen und die in dieser Form erstmalig einem breiteren Publikum vorgestellt werden.

Der dritte Teil beschäftigt sich mit der Umsetzung. Kapitel 6 diskutiert praktische Aspekte bei der Durchführung von Analysen mit dem Rasch Modell und Kapitel 7 beschreibt die konkrete Analyse eines Datensatzes zu Lehrer/inn/enzufriedenheit. Wir haben dabei versucht, jeden Schritt der Analyse nachvollziehbar zu machen. Es werden die Ergebnisse interpretiert und die Analysemethoden verglichen.

Ein weiteres wesentliches Charakteristikum dieses Buches ist die Verwendung von Open-Source-Software, nämlich des freien Programms R für statistische Berechnungen und Grafikerstellung bzw. des Zusatzpakets **eRm**, das Analysen mit dem Rasch Modell ermöglicht. Der Vorteil ist, dass alles kostenlos zur Verfügung steht und dadurch die vorgestellten Methoden unmittelbar am Computer nachvollziehbar werden. Um den Einstieg zu erleichtern, gibt es ein kurzes Tutorial zur Einführung in R im Anhang, in dem die wichtigsten Grundprizipien demonstriert und die Leser/innen von der Installation von R bis hin zum Datenmanagement begleitet werden.

Das Buch ist praxisnahe konzipiert. Steht man am Anfang einer Analyse mit dem Rasch Modell, bereiten manche Aspekte oft Schwierigkeiten. Wie geht man tatsächlich vor? Wie funktionieren die Schätzmethoden und wie sind Ergebnisse zu interpretieren? Was überprüfen die einzelnen Modelltests denn jetzt wirklich? Welche Software kann man zur Analyse verwenden? Hier setzt das Buch an.

Als Zielpublikum haben wir vor allem an Studierende jener Fachrichtungen, in denen empirische Methoden der Datenerhebung relevant sind, gedacht. Als Beispiele seien Psychologie, Pädagogik, Soziologie, Sozial- und Wirtschaftswissenschaften, Medizin oder Sportwissenschaften genannt. Aber auch Praktiker und Wissenschaftler aus solchen Disziplinen sollten den einen oder anderen Aspekt nützlich finden. Statistische Vorkenntnisse sind nicht unbedingt erforderlich, jedoch sind Kenntnisse auf Bachelorniveau von Vorteil. Das Buch versteht sich als Einführungsbuch mit Praxisbezug und als Nachschlagewerk. Es könnte eine einsemestrige Lehrveranstaltung begleiten oder zum Selbststudium dienen.

Auf der Webseite des Verlags ist für dieses Buch ein Downloadbereich eingerichtet. Sie finden diesen auf http://www.utb-shop.de/9783825237868 unter Zusatzmaterialien, von wo Sie den in Kapitel 7 verwendeten Datensatz herunterladen können. Es ist geplant, dort im Laufe der Zeit weitere Materialien und Informationen bereitzustellen.

An dieser Stelle wollen wir uns bei all jenen bedanken, ohne deren Unterstützung dieses Buch nicht zustande gekommen wäre. Michaela Pötscher-Gareiß, Silke Schäfer, Anne Schild, Wolfgang Wiedermann und Elisabeth Zeilinger haben das Buch in seinen unterschiedlichen Phasen begleitet. Sie haben Entwürfe immer wieder gelesen und wichtige Hinweise und Anmerkungen beigesteuert. Besonders bedanken wollen wir uns bei Marco Maier, der wichtige Diskussionspunkte eingebracht und viel Zeit in technische Hilfestellungen und Layout investiert hat. Als Mitentwickler von **eRm** hat er auch zu speziellen Programmentwicklungen beigetragen. Kathrin Gruber hat die komplizierte Objektstruktur grafisch veranschaulicht. Alexandra Grand hat uns den Fragebogen und die Daten, die sie im Zuge ihrer Diplomarbeit erhoben und analysiert hat, zur Verfügung gestellt. Regina Dittrich hat nicht nur Teile gegengelesen, sondern auch als Diskussionspartnerin gedient und Hilfestellung bei so manchen Entscheidungen gegeben. Einige ihrer Studentinnen haben die R-Einführung durchprobiert und wichtiges Feedback gegeben. Mit Norman Verhelst konnten wir einige wichtige Aspekte diskutieren. Schließlich wollen wir uns bei Peter Wittmann, dem Verantwortlichen des Verlags, bedanken. Es ist immer eine Freude sich mit ihm zu unterhalten und mit ihm zusammenzuarbeiten. Seine Kompetenz und Geduld hat uns sehr weitergeholfen.

Ingrid Koller
Rainer Alexandrowicz
Reinhold Hatzinger

Juli, 2012

Unmittelbar nach Fertigstellung dieses Buches ist Reinhold Hatzinger völlig unerwartet aus dem Leben geschieden. Wir trauern um einen großartigen Wissenschaftler, Mentor, Diskussionspartner, Kollegen und Freund. Als herausragender Experte auf seinem Gebiet teilte er sein umfangreiches Wissen mit jeder interessierten Person, hatte für jedes Anliegen stets ein offenes Ohr und förderte junge Wissenschaftler, wann immer er konnte.

Als Mensch und Wissenschaftler hinterlässt er eine große Lücke.

Ihm sei dieses Buch gewidmet.

Ingrid Koller & Rainer Alexandrowicz

Inhaltsverzeichnis

1. Einleitung **1**
1.1. Latente Konstrukte und manifeste Indikatoren 1
1.2. Qualität der Messung 4
1.3. Aufbau des Buchs und weiterführende Literatur 6

2. Das dichotome Rasch Modell **9**
2.1. Zentrale Eigenschaften des Rasch Modells 14
2.1.1. Eindimensionalität – Homogenität der Items 15
2.1.2. Lokale stochastische Unabhängigkeit 16
2.1.3. Spezifische Objektivität/Stichprobenunabhängigkeit . 19
2.1.4. Streng monoton steigende Itemcharakteristikkurven . 22
2.1.5. Suffizienz 25
2.1.6. Zusammenspiel der Eigenschaften 26

3. Schätzung der Modellparameter **29**
3.1. Die Likelihood 30
3.2. Die Joint ML-Schätzung (JML) 35
3.3. Die Conditional ML-Schätzung (CML) 36
3.3.1. Das bedingte Schätzprinzip 37
3.3.2. Die Conditional Likelihood 38
3.4. Normierung der Itemparameter 42
3.5. Schätzung der Personenparameter 42
3.6. Die Präzision der Parameterschätzung 44
3.7. Die Behandlung fehlender Werte 45
3.8. Parameterschätzung mit dem Paket **eRm** 46
3.8.1. Ein Beispielsdatensatz 47
3.8.2. Schätzung der Item- und Personenparameter 48

4. Modellprüfung **61**
4.1. Subgruppeninvarianz 62
4.1.1. Die Wahl des Teilungskriteriums 63
4.1.2. Der Andersen-Likelihood-Ratio-Test 67

4.1.3. Der Wald-Test 77
4.1.4. Grafische Modellkontrollen 79
4.2. Itemhomogenität: Der Martin-Löf-Test 90

5. Nicht-parametrische Überprüfung der Modellgültigkeit 99
5.1. Kleine Stichproben 99
5.2. Exakte und quasi-exakte Tests: Einführung 101
5.2.1. Quasi-exakte Tests beim Rasch Modell 106
5.3. Tests: Lokale stochastische Unabhängigkeit und Homogenität 114
5.3.1. Globale Test-Statistiken 114
5.3.2. Überprüfung auf Itemebene 122
5.4. Tests: Überprüfung der Subgruppeninvarianz 135
5.4.1. Globale Test-Statistik 136
5.4.2. Überprüfung auf Itemebene 142
5.5. Test: Überprüfung unterschiedlicher Itemtrennschärfen . . . 150

6. Praktische Hinweise zur Durchführung einer Itemanalyse 157
6.1. Theoretische Analyse und Ablaufmodell 157
6.2. Alpha-Korrektur und multiples Testen 160
6.3. Eine mögliche Vorgehensweise 167

7. Die Analyse von Items anhand eines realen Datensatzes 171
7.1. Der Fragebogen 171
7.2. A priori Hypothesen 172
7.3. Aufbereitung des Datensatzes 173
7.4. Parametrische Überprüfung des Rasch Modells 175
7.5. Überprüfung des Rasch Modells mit quasi-exakten Tests . . . 197
7.6. Gegenüberstellung der Ergebnisse 215

A. Einstieg in R in drei Sessions 217
A.1. Session 1: R-Basics und die Umgebung 217
A.1.1. Installation von R 217
A.1.2. Einfaches Rechnen 219
A.1.3. Variablen 220
A.1.4. Vektoren und Matrizen 221
A.1.5. Arten von Variablen 225
A.1.6. Funktionen 230
A.1.7. Erweiterungspakete (R-Packages) 231
A.1.8. Die Arbeitsumgebung 233
A.1.9. Speichern von Objekten 234
A.1.10. Zwei Hinweise: Editoren und die R-Hilfe 235

A.2. Session 2: Umgang mit Datensätzen 236
A.2.1. Data Frames 236
A.2.2. Information über den Inhalt von Datensätzen 243
A.3. Session 3: Adaptieren von Datensätzen 247
A.3.1. Fehlende Werte (*missing values*) 248
A.3.2. Auswählen von Teilen eines Datensatzes 251
A.3.3. Erstellen neuer Variablen aus vorhandenen Variablen 252
A.3.4. Umkodieren bei Matrizen und Data Frames 255

B. eRm: Ein R-Package zur Analyse von Rasch Modellen 257

C. Verwendete Symbole 261

Literaturverzeichnis 265

Index 269

1. Einleitung

Manchmal ist das Ziel wissenschaftlicher Untersuchungen herauszufinden, welche Ausprägung eine bestimmte Person bei einem bestimmten Merkmal (z. B. Empathie, Arbeitszufriedenheit) hat und/oder ob daraus eine Vorhersage der Ausprägung bei einem anderen Merkmal (z. B. altruistisches Handeln, Arbeitsmotivation) abgeleitet werden kann. Weiters ist es oft von Interesse mögliche Unterschiede hinsichtlich der Ausprägung zwischen Personengruppen zu untersuchen (z. B. Unterschiede zwischen Geschlechtern, Berufsgruppen oder experimentellen Bedingungen). Von Interesse sind auch Untersuchungen hinsichtlich Veränderungen der Merkmalsausprägung nach wiederholter Erfassung (z. B. Wirksamkeitsstudien).

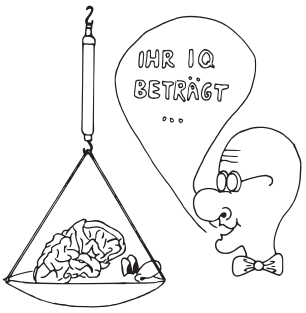
Die zu untersuchenden Objekte können Personen, Tiere oder Gegenstände darstellen, wobei der Fokus in diesem Buch ausschließlich auf Personen gerichtet ist. Die Merkmale können ebenso ganz unterschiedliche Inhalte darstellen, welche aus den unterschiedlichsten Wissenschaftsdisziplinen stammen können. Wir beschäftigen uns in diesem Buch vorwiegend mit psychologischen Merkmalen, wobei wir versuchen die Darstellung so allgemein wie möglich zu halten. Somit sollte der eigentlich methodische Inhalt auf alle möglichen Anwendungsinhalte verallgemeinerbar sein.

1.1. Latente Konstrukte und manifeste Indikatoren

Was sind latente Konstrukte? Latente Konstrukte sind nicht direkt beobachtbare Merkmale. Im psychologischen Kontext sind solche Merkmale wie Leistungsmerkmale (z. B. Intelligenz und Rechtschreibfähigkeit), Persönlichkeitseigenschaften (z. B. Extraversion oder Neurotizimus), Einstellungen (z. B. gegenüber Suizid oder Behinderung), Interessen (z. B. Berufsinteresse oder Freizeitinteresse), Zufriedenheit (z. B. Arbeitszufriedenheit) und vieles mehr.

Diese Merkmale sind nicht direkt messbar und werden in der Fachterminologie als latente Konstrukte bezeichnet. Erfassbar werden diese Merkmale erst durch sogenannte beobachtbare Indikatoren.

Das physikalische Merkmal Körpergewicht können wir mit einer Waage messen. Die Messung des Körpergewichts funktioniert sehr genau. Jedoch für das psychologische Merkmal Intelligenz existiert keine Waage. Die Intelligenz ist ein latentes Konstrukt, für das man nicht einfach eine „Intelligenzwaage" bauen kann, bei der wir dann zur Bestimmung des Intelligenzquotienten (IQ) nur noch unseren Kopf auf die Waage legen müssen.



Die Erfassung von latenten Konstrukten, z. B. im psychologischen Bereich, ist ein viel komplexeres Unterfangen. Hinter jedem latenten Konstrukt steckt ein komplexes Theoriemodell, welches die inhaltliche Bedeutung des zu erfassenden Konstruktes abbilden muss. Dieses Theoriemodell wird zur Generierung von manifesten Indikatoren genutzt. Manifeste Indikatoren können Aufgaben in einem Test, oder Fragen in einem Fragebogen sein (beide Arten werden auch als Items bezeichnet), die die Erfassung der Ausprägung bei einem bestimmten latenten Merkmal ermöglichen sollen. Die Reaktionen bzw. Antworten auf Items soll dann Rückschlüsse auf die Ausprägung eines untersuchten latenten Konstruktes ermöglichen. Grafisch wird das latente Konstrukt als Kreis und die manifesten Indikatoren als Rechtecke dargestellt (siehe Abbildung 1.1). Die Linien zwischen latentem Konstrukt und manifesten Indikatoren stellen die Beziehung zwischen den beiden dar. Sie geben uns an, dass die Beantwortung der manifesten Indikatoren von der Ausprägung eines latenten Konstruktes abhängig ist.

Das theoretische Gerüst eines latenten Konstruktes, wie im unteren Beispiel die Fähigkeit des Grundrechnens, inkludiert nicht nur theoretische Inhalte, sondern auch mögliche Situationen und Komplexitätsstufen, die die Aufgabenschwierigkeit beeinflussen.

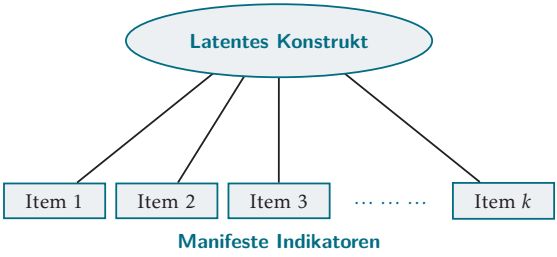


Abbildung 1.1.: Latentes Konstrukt (Kreis) und manifeste Indikatoren (Rechtecke).

Ein Beispiel: Grundrechnen bei Kindern – aller Anfang ist schwer

Zur Erfassung der Fähigkeit Grundrechnen bei Kindern muss zuerst geklärt werden was Grundrechenarten (latentes Konstrukt) sind. Weiters muss man sich überlegen, was sie beinhalten und wie sie erfassbar, also messbar gemacht werden können (Operationalisierung = das latente Konstrukt messbar machen). Dabei soll immer die Zielpopulation (in unserem Fall Kinder in einem bestimmten Alter) berücksichtigt werden. Dieses Beispiel klingt erstmal einfach, ist jedoch wesentlich komplizierter als es erscheint. Allgemein erklärt wird nicht nur zwischen den vier Grundrechenarten (Inhalt) und der Vorgabeart unterschieden. Die Vorgabeart betrifft die Handlungsebene, z. B. die Vorgabe der Rechenaufgaben in arabischer Notation, Einkleidung in Textaufgaben, Berechnung der Aufgabe, Finden von möglichen Fehlern. Weiters kann es noch Komponenten geben, die die Schwierigkeit der Aufgaben beeinflussen: Diese beinhalten unter anderem die Größe der Zahlen, Anordnung der Zahlen (z. B. 25 + 5 vs. 5 + 25, d. h. größere Zahl vorne oder hinten in der Rechenaufgabe), Kombinationen von unterschiedlichen Grundrechenarten, ob Zehnerüberschreitungen vorhanden sind oder nicht und vieles mehr (Komplexität).

Aus diesem theoretischen Gerüst und den Kombinationsmöglichkeiten der Inhalte und Handlungsebenen lassen sich Items, Reize oder Ähnliches mit unterschiedlichen Schwierigkeiten ableiten, durch deren Beantwortung oder Reaktion die Erfassung der Ausprägung eines latenten Konstruktes möglich wird. Die Ausprägung einer Person beim latenten Konstrukt (z. B. der Grad ihrer Intelligenz) ist durch die Personeneigenschaft oder Personen-

fähigkeit festgelegt. Sind die Fähigkeitsausprägung einer Person und die Schwierigkeit eines Items bekannt, können wir wieder Rückschlüsse auf das Antwortverhalten bei dem vorgelegten Item ziehen. Die Items werden in der Fachterminologie als manifeste Indikatoren bezeichnet und sind direkt erfassbar bzw. beobachtbar. Dabei wird die Schwierigkeit eines Items für die Zielpopulation als Itemschwierigkeit bezeichnet.

1.2. Qualität der Messung

Genauso wie das Körpergewicht mit einer Waage messbar ist, soll auch die Ausprägung des latenten Konstruktes messbar sein. Die Messung mit der Waage ist, unter Voraussetzung der richtigen Eichung, als genau einzustufen. Anstatt einer Waage generieren wir Aufgaben oder Items, durch deren Beantwortung wir Rückschlüsse auf die Ausprägung des latenten Konstruktes ziehen können. Dies bedeutet, dass wir die Ausprägung der Personen messbar machen wollen. Somit besteht unser Messinstrument aus den Aufgaben oder Items, die nicht nur inhaltlich gültig sein müssen (theoriegeleitete Konstruktion), sondern auch weiteren messtheoretischen Qualitätskriterien entsprechen sollen. Sind diese Kriterien erfüllt, kann von einer hinreichend genauen Messung gesprochen werden.

Eine erfolgreiche Fragebogen- oder Testkonstruktion inkludiert eine Überprüfung der Qualität der Items anhand einer zuvor gut gewählten Analysestichprobe. Diese soll bereits die eigentliche Zielpopulation, für die der Test oder der Fragebogen bestimmt ist, gut abbilden. Zur Vereinfachung sprechen wir zukünftig nur noch von Tests. Damit sind z. B. Einstellungstests, Wissenstests, Persönlichkeitsfragebögen und vieles mehr gemeint, welche man allgemein als Skalen bezeichnet. Es gilt dann anhand der Daten zu überprüfen, ob die Qualität der Items den testtheoretischen Anforderungen genügt und sich das Instrument somit für die Untersuchung der Fragestellungen eignet. Im Folgenden lernen wir wichtige Qualitätskriterien kennen. Die kurzen Erklärungen dienen dabei nur als Einführung, die genaue Beschreibung erfolgt in Kapitel 2.1.

Eindimensionalität

Eine wichtige Anforderung an die Items ist, dass die Beantwortung der Items vorrangig auf die Fähigkeitsausprägung eines untersuchten latenten Konstruktes zurückführbar ist. Das bedeutet, dass alle Items dieselbe Dimension (Eindimensionalität bzw. Homogenität der Items) erfassen müssen.

Auf Personenseite sollten die Items für alle inhaltlich relevanten Personengruppen (z. B. Geschlecht, Alter und Kultur) gleich schwierig sein und somit dieselbe Schwierigkeitsanordnung im Test aufweisen (Invarianz). Diese Forderung wird in der Fachterminologie als die Überprüfung von Differential Item Functioning (DIF) bezeichnet. Liegt in keiner der untersuchten Personengruppen DIF für Items vor, dann funktionieren die Items in allen untersuchten Personengruppen gleich. Gilt diese Eigenschaft für einen Test über alle untersuchten Personengruppen, so wird keine dieser Personengruppen benachteiligt und der Test stellt ein faires Instrument dar.

Trennschärfe

Ein weiteres sehr wichtiges Qualitätskriterium ist die Trennschärfe der einzelnen Items. Ein Item mit einer guten Trennschärfe kann zwischen fähigeren und weniger fähigen Personen gut unterscheiden (diskriminieren). Das stellt in der Praxis ein wichtiges Qualitätskriterium dar, da man mit Items, die eine zu geringe Trennschärfe aufweisen, keine Entscheidung hinsichtlich der Fähigkeitsausprägung treffen kann. Wenn z. B. ein Rechentest nicht zwischen Kindern mit mathematischen Schwächen und Kindern ohne mathematische Schwächen unterscheiden kann, dann besteht auch nicht die Möglichkeit erstere zu identifizieren und diesen in weiterer Folge eine geeignete Förderung zukommen zu lassen.

Abhängigkeiten zwischen Items

Des Weiteren sollen die Items in einem Test nicht zu stark voneinander abhängig sein. Die Wahrscheinlichkeit ein Item zu lösen, soll nur von dem Item selbst und nicht von der Lösungswahrscheinlichkeit eines anderen Items abhängen. Der Grund einer Verletzung könnte hier eine zu große Ähnlichkeit zwischen Items sein, oder aber dass die Inhalte von Items aufeinander aufbauen. Bei Vorgabe von Items, die sich inhaltlich sehr ähnlich sind, oder denselben Inhalt erfragen, würde man den zu testenden Personen mehr Items als notwendig vorlegen. Diese voneinander zu stark abhängigen Items erbringen im Weiteren wenig zusätzliche Information zur Bestimmung der Ausprägung eines latenten Konstruktes. Die Vorgabe dieser Items würde zu einer unnötigen Testverlängerung führen. Weiters könnte die Vorgabe von zu stark abhängigen Items wieder zu einer Benachteiligung von Personengruppen führen. Stellen wir uns vor, dass das Lösen eines Items die Lösungswahrscheinlichkeit eines nachfolgenden Items positiv beeinflusst (z. B. durch Lernen), so bedeutet das, dass die Wahrscheinlichkeit das nachfolgende Item zu lösen steigt. Diese Aussage gilt jedoch nicht, wenn das

vorherige Item nicht gelöst wurde. Somit hat die eine Gruppe (vorheriges Item gelöst) einen Vorteil gegenüber der anderen Gruppe (vorheriges Item nicht gelöst).

Die Randsummen: Ausreichende Information über die Personenfähigkeit und die Itemschwierigkeit

Wenn diese und noch weitere Eigenschaften, die wir in Kapitel 2.1 genauer betrachten werden, zutreffen, dann enthalten die Anzahl der gelösten (zugestimmten) Items einer Person in einem Test oder die Anzahl der positiven (zugestimmten) Antworten für ein Item alle Informationen für die Personenfähigkeit und die Itemschwierigkeit. Dann ist es irrelevant, welche Items gelöst wurden, und von welchen Personen das Item gelöst wurde (siehe Kapitel 2.1.5). Diese Eigenschaft hat einen zentralen Stellenwert, denn erst wenn diese zutrifft, ist das Aufsummieren der z. B. richtig gelösten Aufgaben einer Person zu einer Randsumme und die weitere Analyse derer gerechtfertigt.

Dichotome Items

In diesem Buch beschäftigen wir uns mit dem Typ der dichotomen Items (auch zweikategoriale oder binäre Items genannt). Das sind Items mit zwei möglichen Antwortalternativen. Dichotom kodierte Items in einem Leistungstest bedeuten z. B. gelöst (kodiert mit 1) oder nicht gelöst (kodiert mit 0) oder in einem Einstellungsfragebogen z. B. stimme zu (1) oder stimme nicht zu (0). Dieser Itemtypus kann mit dem Rasch Modell (siehe z. B. Fischer und Molenaar, 1995) analysiert werden, welches sich zur Überprüfung der vorher genannten Qualitätskriterien besonders gut eignet.

1.3. Aufbau des Buchs und weiterführende Literatur

Aufbau des Buchs

Das Buch beinhaltet nicht nur eine kurze theoretische Einführung zum Rasch Modell im Allgemeinen, sondern ebenso eine praktische Einführung. Diese betrifft das in R (R Development Core Team, 2012) implementierte Paket **eRm** (Mair et al., 2012), welches sich für die Analyse von Items, also für die Überprüfung auf Geltung des Rasch Modells und somit Zutreffen der geforderten Qualitätskriterien, hervorragend eignet.

Das Buch gliedert sich in drei Teile. Im ersten Teil (Kapitel 1 und 2) werden die Grundideen der Qualitätskontrolle bei Messungen latenter Variablen und ihrer formalen Umsetzung im Rasch Modell vorgestellt. Der zweite Teil beschreibt in Kapitel 3 allgemeine Prinzipien der Maximum-Likelihood-Methode sowie ihrer konkreten Anwendung zur Schätzung der Modellparameter im Rasch Modell. Hier wird auch auf fehlende Werte eingegangen. In Kapitel 4 werden grafische Methoden und parametrische Tests zur Überprüfung der Gültigkeit des Rasch Modells vorgestellt. Neueste Entwicklungen zur Modellüberprüfung, sogenannte quasi-exakte Tests, die sich besonders bei kleinen Stichproben eignen, sind das Thema von Kapitel 5. Die drei Kapitel des zweiten Teils sind so aufgebaut, dass zunächst kurze theoretischen Einführungen zu einzelnen Aspekten der Analyse mit dem Rasch Modell gegeben werden. Im Anschluß wird die jeweils dazugehörige computergestützte Umsetzung in **eRm** präsentiert. Dabei werden die dazu notwendigen R-Befehle (Eingaben für die Berechnung) erklärt. Die verwendeten Daten sind direkt in R verfügbar. Den Leser/inne/n soll so ermöglicht werden, jeden einzelnen Schritt selbstständig nachzuvollziehen und das Gelernte direkt praktisch anwenden zu können.

Der dritte Teil betrifft die praktische Anwendung. Kapitel 6 diskutiert mögliche Vorgehensweisen bei der Analyse von Daten mit dem Rasch Modell. Kapitel 7 präsentiert beispielhaft eine konkrete Analyse anhand der im Buch behandelten Methoden und Interpretationen der Ergebnisse. Schließlich folgt ein kurze Einführung in R, die Anfänger/inn/en die wichtigsten Grundprinzipien der Bedienung der Software nahebringen soll.

Das Buch soll einerseits die selbstständige Anwendung des Rasch Modells ermöglichen, als auch als Nachschlagewerk dienen.

Weiterführende Literatur

An dieser Stelle möchten wir Ihnen einige Literaturempfehlungen für Item Response Modelle und für das Programm R mit auf den Weg geben.

Als klassisches Lehrbuch für Item Response Theorie ist Rost (2004) anzuführen. Das deutschsprachige Buch beinhaltet eine gute Einführung in Modelle der Item Response Theorie. Es ist sowohl für Anfänger als auch als Nachschlagewerk geeignet. Außerdem inkludiert es eine CD mit Übungen, u. a. auch mit Beispielen für die Anwendung der Programme WINMIRA und MULTIRA.

Ein weiteres deutschsprachiges Buch ist von Strobl (2010). Sie gibt eine kurze Einführung in das Rasch Modell. Es enthält auch neue Modelltests sowie die Einführung in zwei R-Pakete zur Anwendung von Item Response Modellen.

Als englischsprachiges Lehrbuch kann jenes von de Ayala (2009) empfohlen werden, in welchem u. a. auch die Beschreibungen zu den kommerziellen Programme BIGSTEPS, BILOG, MULTILOG und SAS enthalten sind.

Ein anspruchsvolleres Buch ist das englischsprachige Werk von Fischer und Molenaar (1995). In diesem Buch sind die theoretischen Grundlagen zu unterschiedlichen Modellen der Familie der Rasch Modelle zu finden. Wenn man sich für das Rasch Modell interessiert, darf es in der Bibliothek nicht fehlen.

Weiters ist zu erwähnen, dass ein umfassendes englischsprachiges Werk zu diesem Thema geplant ist und Ende 2012 oder zu Beginn 2013 erscheinen wird. Das *Handbook of IRT* von van der Linden und Hambleton (2013) wird 3 Bände umfassen. In diesen Bänden sind einerseits die theoretischen Grundlagen von Item Response Modellen sowie Programme für die Anwendung dieser Modelle dargestellt.

Wir arbeiten in unserem Buch mit dem Open-Source-Programm R. Wir möchten dazu zwei Literaturempfehlungen geben.

Alexandrowicz (2012) gibt eine Einführung in R in 10 Schritten. Neben der allgemeinen Einführung behandelt das Buch noch einige statistische Auswertungsmethoden sowie Grundlagen der Programmierung in R. Auch dieses Werk wird in Bälde erscheinen.

Schließlich wollen wir noch das Buch von Hatzinger et al. (2011) empfehlen. Neben einer ausführlichen praxisnahen Einführung in R werden grundlegende (aber zum Teil auch fortgeschrittenere) statistische Methoden zur angewandten Datenanalyse behandelt.

2. Das dichotome Rasch Modell. Eine theoretische Einführung

Das Rasch Modell für dichotome Daten wird durch die Modellgleichung (2.1) beschrieben:

$$P(X_{vi} = x_{vi} \,|\, \theta_v, \beta_i) = \frac{\exp[x_{vi}(\theta_v - \beta_i)]}{1 + \exp(\theta_v - \beta_i)} = p_{vi} \tag{2.1}$$

- θ_v: Personenfähigkeitsparameter für $v = 1, \ldots, n$
- β_i: Itemschwierigkeitsparameter für $i = 1, \ldots, k$
- X_{vi}: Variable für eine Antwort einer Person v auf ein Item i
- x_{vi}: konkreter Wert, den die Variable X_{vi} annimmt (0 oder 1)

Die Gleichung drückt die Wahrscheinlichkeit für eine Antwort x_{vi}, gegeben (symbolisiert durch $|$) der Personenfähigkeit θ_v und der Itemschwierigkeit β_i aus. Eine kürzere Schreibweise wäre der Ausdruck p_{vi}, welcher auf der rechten Seite dargestellt ist. Wenn wir im Buch auf die Gleichung zurückkommen, wird zukünftig nur noch dieser Ausdruck benutzt. Bevor wir uns jedoch der Gleichung etwas genauer zuwenden, müssen wir ein paar grundsätzliche Ausdrücke kennenlernen.

Stellen wir uns vor, wir haben einer bestimmte Anzahl von Personen eine bestimmte Anzahl von Items vorgegeben. Betrachten wir eine bestimmte Person aus unserer Stichprobe, dann handelt es sich um die v-te Person der gesamten Anzahl an n Personen. Somit haben wir insgesamt $v = 1, \ldots, n$ Personen getestet. Betrachten wir ein bestimmtes Item, dann nennen wir es das Item i, wobei wir $i = 1, \ldots, k$ Items vorgegeben haben. Somit stellen die beiden Indizes v für die Person und i für das Item immer bestimmte Personen oder bestimmte Items in unserem Datensatz dar, welcher eine Größe (Dimension) von $n \times k$ aufweist.

Der Buchstabe n steht für die Anzahl von Personen, also somit der Anzahl von Zeilen im Datensatz. Der Buchstabe k bezeichnet die Anzahl der Items, also

die Anzahl der Spalten im Datensatz. Werden diese beiden Zahlen miteinander multipliziert, dann erhält man die Anzahl von Zellen des Datensatzes, dessen Darstellung in der Fachterminologie als Datenmatrix bezeichnet wird. Abbildung 2.1 zeigt eine allgemeine Darstellung einer solchen Datenmatrix.

Die x_{vi}, auf die wir gleich noch näher eingehen, stellen die einzelnen Einträge in der Matrix dar. Die Auslassungspunkte in der Darstellung geben an, dass hier noch eine unbestimmte Anzahl an Personen und Items vorhanden ist. Somit können wir die eigentliche Größe der Datenmatrix nicht genau bestimmen und sie nur als $n \times k$ beschreiben.

Die Einträge haben zwei Indizes, nämlich v für Personen und i für Items. Der erste Index stellt die fortlaufende Nummer der Personen (Zeilen) dar, welcher von $1 \ldots, n$ geht. Der zweite Index bezeichnet die fortlaufende Nummer der Items (Spalten) und geht von $1 \ldots, k$. Der letzte Eintrag in der Datenmatrix ist x_{nk} und stellt den Eintrag der letzten Person und des letzten Items dar.

$$\begin{array}{cccccc}
x_{11} & x_{12} & \cdots & x_{1i} & \cdots & x_{1k} \\
x_{21} & x_{22} & \cdots & \cdots & \cdots & x_{2k} \\
x_{31} & x_{32} & \cdots & \cdots & \cdots & x_{3k} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{v1} & x_{v2} & \cdots & x_{vi} & \cdots & x_{vk} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{ni} & \cdots & x_{nk}
\end{array}$$

Abbildung 2.1.: Datenmatrix

Wenn wir hypothetisch annehmen, dass wir k Items n Personen vorgeben, dann stellt die Zufallsvariable X_{vi} das noch unbekannte Antwortverhalten der Person dar. Vor der Vorgabe des Items, wissen wir noch nicht, ob die Person das Item löst oder nicht. Wie die Person auf ein Item antwortet, also der tatsächliche Ausgang der Aufgabenbearbeitung (der in unserem Fall nur 0 oder 1 darstellen kann) wird mit x_{vi} bezeichnet. Somit ist das allgemeine Antwortverhalten einer Person auf ein Item mit $X_{vi} = x_{vi}$ beschrieben, wobei x_{vi} bei der Lösung/Zustimmung des Items den Wert 1 und bei Nichtlösung/keine Zustimmung des Items den Wert 0 annimmt.

Wenn wir nochmals die Gleichung (2.1) betrachten, dann sehen wir, dass die linke Seite der Gleichung genau die gerade erklärte allgemeine Schreibweise

des Antwortverhaltens einer bestimmten Person auf ein bestimmtes Item darstellt.

In der Praxis haben wir es jedoch mit realen Datenmatrizen zu tun, in der statt dem Eintrag x_{vi}, in jeder Zelle die Antwortmöglichkeit 0 oder 1 steht. In der folgend dargestellten Datenmatrix in Abbildung 2.2 sind Antworten von fünf Personen auf vier Items dargestellt. Damit beinhaltet die Matrix $5 \times 4 = 20$ Einträge.

Rechts außen sind mit r_v die Personenscores r für jede einzelne Person v, also die Summe der Einträge einer jeden Zeile, dargestellt. Unterhalb sind mit c_i die Itemscores c für jedes Item i, also die Summe der Einträge einer jeden Spalte, dargestellt. In der klassischen Testtheorie sind der Personenscore als

	I1	I2	I3	I4	Summe r_v
A	1	0	1	1	3
B	0	0	0	0	0
C	0	1	1	0	2
D	1	1	1	1	4
E	0	1	1	0	2
Summe c_i	2	3	4	2	–

Abbildung 2.2.: Datenmatrix für fünf Personen und vier Items

die Personenfähigkeit und der Itemscore als die Itemleichtigkeit definiert. Es gilt jedoch zu beachten, dass die Berechnung eines Personenscores und dem Itemscores nur zulässig ist, wenn für die Daten bestimmte Anforderungen an eine Messung gelten. Diese werden wir in diesem Kapitel noch genauer kennenlernen.

Im Rasch Modell werden Personenfähigkeit und Itemschwierigkeit anhand der beobachteten Daten geschätzt, wobei die unbekannten zu schätzenden Parameter für die Personenfähigkeiten mit $\hat{\theta}_v$ und für die Itemschwierigkeit mit $\hat{\beta}_i$ bezeichnet werden. Das Dach über θ und über β bedeutet, dass es sich hierbei nicht um die wahren Parameter, sondern um die Schätzwerte für diese Parameter handelt.

Die zentrale Frage ist: Wie hoch ist die Wahrscheinlichkeit p_{vi} (p kommt aus dem Englischen und steht für probability) einer Person v mit der geschätzten Personenfähigkeit θ_v auf ein Item i mit der geschätzten Itemschwierigkeit β_i mit x_{vi} zu antworten. Diese Frage ist, wie bereits erklärt, auf der linken Seite der Gleichung (2.1) formuliert.

Möchte man die Lösungswahrscheinlichkeit berechnen, dann wird auf der linken Seite $X_{vi} = 1$ (d. h. das Item wurde gelöst), wie in der Gleichung (2.2), geschrieben. Für die Gegenwahrscheinlichkeit wird dann, wie in Gleichung (2.3), $X_{vi} = 0$ (d. h. das Item wurde nicht gelöst) geschrieben.

Das x_{vi} spielt ebenso auf der rechten Seite der Gleichung (2.1), genauer gesagt im Zähler, eine bedeutsame Rolle. Wird das Item gelöst, wird der Ausdruck $\theta_v - \beta_i$ im Zähler mit 1 multipliziert und führt zu Gleichung (2.2). Wird das Item nicht gelöst, wird der Ausdruck im Zähler mit 0 multipliziert und ergibt den Zähler der Gleichung (2.3). Warum das Ergebnis 1 lautet, wird im nächsten Absatz erklärt.

$$P(X_{vi} = 1 \,|\, \theta_v, \beta_i) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)} \tag{2.2}$$

$$P(X_{vi} = 0 \,|\, \theta_v, \beta_i) = \frac{1}{1 + \exp(\theta_v - \beta_i)} \tag{2.3}$$

Ein wichtiger Bestandteil der Gleichung ist die Abkürzung $\exp(\cdot)$. Diese Abkürzung stellt die Exponentialfunktion dar, welche die Eulersche Zahl e ($e = 2{,}71828\ldots$) als Basis hat. Dabei ist $\exp(x)$ die Exponentialfunktion einer Variable x. Ist nun der Ausgang $X_{vi} = 0$ von Interesse, dann wird der Zähler der Gleichung (2.1) zu $\exp(0) = 1$.

Wir benutzen in diesem Buch die Schreibweise $\exp(\cdot)$. Eine alternative Schreibweise ist e^x, wobei damit dieselbe Funktion gemeint ist.

Nebenbei anzumerken wäre noch, dass die Umkehrfunktion von $\exp(\cdot)$ die natürliche Logarithmusfunktion $\ln(\cdot)$ darstellt (im Englischen auch mit $\log(\cdot)$ bezeichnet). Zum Beispiel ergibt $\exp(0) = 1$ und $\ln(1) = 0$. Allgemein formuliert: $\exp(a) = b$ und $\ln(b) = a$.

Die Exponentialfunktion $\exp(\cdot)$ benötigen wir nicht nur zur Berechnung der Wahrscheinlichkeit, sondern sie spielt eine zentrale Rolle beim Verlauf der Itemcharakteristikkurve, die wir später in diesem Kapitel noch kennen lernen werden.

Kommen wir zurück zum Ausdruck $\theta_v - \beta_i$. Dieser Teil der Gleichung stellt die Differenz zwischen Personenfähigkeit und Itemschwierigkeit dar. Ist die Person fähiger als das Item schwierig ist, dann wird das Ergebnis im positiven Bereich liegen. Ist das Item schwieriger als die Person fähig ist, so wird die

Differenz im negativen Bereich sein. Ist die Person jedoch genau so fähig wie das Item schwierig ist, dann ergibt sich eine Differenz von 0.

Nachdem wir alle wichtigen Bestandteile der Gleichung kennen, können wir die Lösungswahrscheinlichkeit p_{vi} berechnen. Wie in den Grundlagen der Wahrscheinlichkeitstheorie, berechnet man die Wahrscheinlichkeit eines Ereignisses über die Anzahl der günstigen Ereignisse dividiert durch die Anzahl der möglichen Ereignisse. Im Falle von dichotomen Items gibt es nur die Ereignisse 0 und 1. Somit erhalten wir im Nenner der Gleichung $(\exp[0 \cdot (\theta_v - \beta_i)]) + (\exp[1 \cdot (\theta_v - \beta_i)])$. Der erste Teil wird mit 0 multipliziert und ergibt $\exp(0) = 1$. Somit lautet der Nenner der Gleichung $1 + \exp(\theta_v - \beta_i)$.

Die Itemcharakteristikkurve

Zum Schluss kommen wir nocheinmal zurück zu der Exponentialfunktion. Die Exponentialfunktion $\exp(\cdot)$ bestimmt den Verlauf der Itemcharakteristikkurve (ICC; engl.: item characteristic curve), welche die Lösungswahrscheinlichkeit für ein Item in Abhängigkeit der Personenfähigkeit und der Itemschwierigkeit abbildet (siehe Abbildung 2.3). Auf der y-Achse wird die Wahrscheinlichkeit ein Item zu lösen aufgetragen, wobei dieser Bereich von 0 bis 1 geht. Auf der x-Achse sind die möglichen Personenfähigkeiten dargestellt, wobei diese immer von minus Unendlich ($-\infty$) bis plus Unendlich ($+\infty$) verlaufen. In Abbildung 2.3 sehen wir die Kurve für ein Item mit $\beta_i = 1.1$. Die Charakteristik der Kurve zeigt, dass bei konstanter Itemschwierigkeit die Lösungswahrscheinlichkeit eines Items mit steigender Personenfähigkeit ansteigt.

Nehmen wir also an, eine Person ist genauso fähig wie das Item schwierig ist. In diesem Fall ergibt die Differenz von $\theta_v - \beta_i = 0$. Damit wird der Zähler $\exp(0) = 1$. An dieser Stelle beträgt die Diskrimination (siehe Seite 5) des Items (bzw. die Differenzierungsleistung eines Items für unterschiedliche Personenfähigkeiten) definitionsgemäß 1. Berechnet man nun die Lösungswahrscheinlichkeit des Items, dann liegt diese genau bei 0.5 (50%).

Laut Konvention im Rasch Modell wird die Itemschwierigkeit für ein bestimmtes Item durch den Punkt, an dem die Lösungswahrscheinlichkeit des Items 50% beträgt und die ICC ihren steilsten Anstieg hat, definiert. Somit sieht man, dass auf der x-Achse nicht nur das mögliche Kontinuum der Personenfähigkeit, sondern auch die Itemschwierigkeit abgebildet ist.

Um das zu verdeutlichen, stellen wir uns ein Item mit der geschätzten Itemschwierigkeit von $\beta_i = 1$ und drei Personen mit den Personenfähigkeitsparametern $\theta_1 = 0$, $\theta_2 = 1$ sowie $\theta_3 = 2$ vor. Setzen wir (probieren Sie es mit