psychlehrbuchplus
hogrefe.de

Michael Eid · Katharina Schmidt

Testtheorie und Testkonstruktion

BACHELORSTUDIUM PSYCHOLOGIE

HOGREFE

Testtheorie und Testkonstruktion

Bachelorstudium Psychologie

Testtheorie und Testkonstruktion

von Prof. Dr. Michael Eid und Katharina Schmidt

Herausgeber der Reihe:
Prof. Dr. Eva Bamberg, Prof. Dr. Hans-Werner Bierhoff,
Prof. Dr. Alexander Grob, Prof. Dr. Franz Petermann

Testtheorie und Testkonstruktion

von
Michael Eid
und Katharina Schmidt

HOGREFE
GÖTTINGEN · BERN · WIEN · PARIS · OXFORD · PRAG
TORONTO · BOSTON · AMSTERDAM · KOPENHAGEN
STOCKHOLM · FLORENZ · HELSINKI

Prof. Dr. Michael Eid, geb. 1963. Seit 2006 Professor für Methoden und Evaluation an der Freien Universität Berlin.

Dipl.-Psych. Katharina Schmidt, geb. 1963. Seit 2006 wissenschaftliche Mitarbeiterin im Arbeitsbereich für psychologische Diagnostik und Differentielle und Persönlichkeitspsychologie an der Freien Universität Berlin.

Informationen und Zusatzmaterialien zu diesem Buch finden Sie unter
www.hogrefe.de/buecher/lehrbuecher/psychlehrbuchplus

Göttingen · Bern · Wien · Paris · Oxford · Prag · Toronto · Boston
Amsterdam · Kopenhagen · Stockholm · Florenz · Helsinki
Merkelstraße 3, 37085 Göttingen

http://www.hogrefe.de
Aktuelle Informationen · Weitere Titel zum Thema · Ergänzende Materialien

---

Umschlagabbildung: © Digitalstock – L. Halbauer
Satz: ARThür Grafik-Design & Kunst, Weimar
Format: PDF

ISBN 978-3-8409-2161-2

Nutzungsbedingungen:

Der Erwerber erhält ein einfaches und nicht übertragbares Nutzungsrecht, das ihn zum privaten Gebrauch des E-Books und all der dazugehörigen Dateien berechtigt.

Der Inhalt dieses E-Books darf von dem Kunden vorbehaltlich abweichender zwingender gesetzlicher Regeln weder inhaltlich noch redaktionell verändert werden. Insbesondere darf er Urheberrechtsvermerke, Markenzeichen, digitale Wasserzeichen und andere Rechtsvorbehalte im abgerufenen Inhalt nicht entfernen.

Der Nutzer ist nicht berechtigt, das E-Book – auch nicht auszugsweise – anderen Personen zugänglich zu machen, insbesondere es weiterzuleiten, zu verleihen oder zu vermieten.

Das entgeltliche oder unentgeltliche Einstellen des E-Books ins Internet oder in andere Netzwerke, der Weiterverkauf und/oder jede Art der Nutzung zu kommerziellen Zwecken sind nicht zulässig.

Das Anfertigen von Vervielfältigungen, das Ausdrucken oder Speichern auf anderen Wiedergabegeräten ist nur für den persönlichen Gebrauch gestattet. Dritten darf dadurch kein Zugang ermöglicht werden.

Die Übernahme des gesamten E-Books in eine eigene Print- und/oder Online-Publikation ist nicht gestattet. Die Inhalte des E-Books dürfen nur zu privaten Zwecken und nur auszugsweise kopiert werden.

Diese Bestimmungen gelten gegebenenfalls auch für zum E-Book gehörende Audiodateien.

Anmerkung:

Sofern der Printausgabe eine CD-ROM beigefügt ist, sind die Materialien/Arbeitsblätter, die sich darauf befinden, bereits Bestandteil dieses E-Books.

Unseren Kindern
Joshua und Johanna gewidmet

Inhaltsverzeichnis

Vorwort

That this text presents to the reader more problems than it solves is perhaps merely a sign of the youth and vitality of a movement which I believe is destined to revolutionize the human relationship problems of society.

Kelley (1927, S. iv)

Über achtzig Jahre sind vergangen, seit Truman Lee Kelley mit diesem Satz das Vorwort zu einem der ersten Lehrbücher der Testtheorie *(Interpretations of educational measurements)* beendet hat. Über die letzten achtzig Jahre hinweg hat sich die Testtheorie rasant entwickelt und ist zu einem zentralen Methodenfach der Psychologie geworden, das unser Verständnis von den messtheoretischen Grundlagen der empirischen Psychologie entscheidend gefördert hat. Ob sich dadurch die „human relationship problems of society" revolutioniert haben, sei dahingestellt, aber ohne ein fundiertes testtheoretisches Wissen kann man viele Tätigkeiten in der Psychologie nicht adäquat ausfüllen, nicht nur in der Forschung, sondern vor allem auch in der psychodiagnostischen Praxis.

Im Laufe ihrer Entwicklung hat sich die Testtheorie zunehmend ausdifferenziert, es wurden aber auch vielfältige Bezüge zwischen einzelnen Testtheorien deutlich, die es heute ermöglichen, verschiedene testtheoretische Ansätze in einem integrativen Rahmen darzustellen. Dies hilft, Bezüge zu erkennen und „Mythen" zu entlarven. Das vorliegende Lehrbuch basiert auf einer solch integrativen Sicht und stellt die verschiedenen testtheoretischen Ansätze gegliedert nach der Art der Antwortvariablen, die untersucht werden sollen, dar. Klassische Testtheorie und Item-Response-Theorie werden somit integrativ im Rahmen einer gemeinsamen Darstellungsform vermittelt, die an Mellenberghs (1994b) Konzept einer *Generalisierten Linearen Item-Response-Theorie* orientiert ist. Dies steht im Gegensatz zu der weithin getrennten Vermittlung der Klassischen Testtheorie und der Item-Response-Theorie, die eher daran orientiert war, die Unterschiede zwischen diesen Ansätzen herauszuarbeiten, ohne die vielen Gemeinsamkeiten zu sehen. Durch die integrative Darstellung soll das Lernen erleichtert werden und gleichsam die Klassische Testtheorie als eine moderne Testtheorie vermittelt werden, die der Item-Response-Theorie in nichts nachsteht.

Das Buch führt in die Grundlagen der Testtheorie und Testkonstruktion ein. Es wurde Wert darauf gelegt, dieses zentrale Fach der Psychologie verständlich und anwendungsorientiert zu vermitteln, ohne die notwendige fachliche Tiefe zu vernachlässigen. Es richtet sich an Studierende der Psychologie, die ein fundiertes

Verständnis der modernen testtheoretischen Ansätze und ihrer zentralen anwendungspraktischen Implikationen erwerben wollen. Hierzu wird auf reale Anwendungsbeispiele der Psychologie zurückgegriffen, die ausführlich dargestellt werden. Auf einer begleitenden Internetseite (www.hogrefe.de/buecher/lehrbuecher/psychlehrbuchplus) werden alle Datensätze zur Verfügung gestellt, sodass Studierende alle Beispiele nachrechnen können. Um dies einfach zu ermöglichen, werden kommentierte Inputs und Outputs zur Verfügung gestellt, die sich vor allem auf das im Internet frei verfügbare Computerprogramm R beziehen (http://www.r-project.org/). Dies soll es Studierenden ermöglichen, nicht nur Wissen, sondern auch anwendungspraktische Kompetenzen zu erwerben. Auf der Internetseite sind auch Antworten zu den Lernkontrollfragen des Buches zu finden. Das Buch richtet sich aber nicht nur an Studierende, sondern an all jene, die sich ein grundlegendes Verständnis der testtheoretischen Grundlagen und ihrer anwendungspraktischen Bezüge erwerben wollen.

Das Buch ist ein einführendes Buch und daher zwangsläufig in der Breite der behandelten Themen beschränkt. Im Mittelpunkt steht die Vermittlung der grundlegenden Prinzipien anhand der einfachsten Modelle der Testtheorie, die zur Erfassung eindimensionaler Merkmale (Konstrukte) entwickelt wurden. Es gliedert sich in acht Kapitel.

Kapitel 1 führt in die Grundfragen der Testtheorie ein: Warum ist dieses Gebiet für die Psychologie von grundlegender Bedeutung? Worin unterscheidet sich Messen in der Psychologie von der Messung in anderen Naturwissenschaften wie z. B. der Physik? Warum braucht die Psychologie eine eigene Testtheorie und wieso sollte man sich mit dieser Thematik überhaupt beschäftigen? Was versteht man unter einem psychologischen Test und was unter einem psychologischen Konstrukt?

In *Kapitel 2* werden die zentralen Gütekriterien psychologischer Tests und die einzelnen Schritte der Testkonstruktion behandelt. Woran erkennt man einen guten psychometrischen Test? Wie geht man bei der Testkonstruktion vor? Worin unterscheiden sich verschiedene Prinzipien der Testkonstruktion? Wie kann man die Qualität psychologischer Tests sicherstellen?

Kapitel 3 stellt verschiedene Item- und Antwortformate vor. Was muss man bei der Formulierung einer Frage eines Fragebogens beachten? Wie viele Antwortalternativen sollte man in einem Multiple-Choice-Test vorgeben? Wie sollte man Antwortformate benennen?

Die *Kapitel 4 bis 6* behandeln Methoden zur Auswertung von Items eines Tests. Die Kapitel unterscheiden sich in der Art der behandelten Variablen, folgen aber demselben Aufbau. Zentrale Fragen sind: Wie lassen sich Merkmalsunterschiede

zwischen Personen beschreiben? Wie kann man Zusammenhänge zwischen verschiedenen Items erfassen? Wie lassen sich die Ausprägungen von Personenmerkmalen, die nicht direkt beobachtet werden können, schätzen? Was versteht man unter psychometrischen Modellen? Auf welchen Annahmen basieren diese und wie lässt sich deren Gültigkeit überprüfen? *Kapitel 4* widmet sich dichotomen Antwortvariablen und stellt insbesondere das Rasch-Modell und seine Erweiterungen vor. In *Kapitel 5* werden Modelle für Items mit geordneten Antwortkategorien behandelt, insbesondere das Partial-Credit-Modell und seine Spezialfälle sowie Erweiterungen. *Kapitel 6* hat schließlich metrische Antwortvariablen zum Gegenstand, es werden die Grundprinzipien und die verschiedenen Modelle der Klassischen Testtheorie behandelt.

Kapitel 7 überträgt die Grundideen testtheoretischer Modelle auf mehrdimensionale Modelle und stellt deren Grundideen exemplarisch dar. Worin unterscheiden sich eindimensionale von mehrdimensionalen Modellen? Was sind die Grundideen der konfirmatorischen und der exploratorischen Faktorenanalyse?

In *Kapitel 8* wird schließlich mit der Normierung ein zentrales Problem der psychologischen Diagnostik behandelt. Was bedeutet ein Testwert? Wie lässt er sich interpretieren? Wie kann man feststellen, ob sich zwei Personen unterscheiden oder eine Person sich verändert hat?

Eine zusätzliche *Formelsammlung*, in der die Formeln zum schnellen Nachschlagen zusammengestellt sind, findet sich auf der Internetseite www.hogrefe.de/buecher/lehrbuecher/psychlehrbuchplus.

Das Buch wurde so konzipiert, dass die Themen im Rahmen einer einsemestrigen Lehrveranstaltung behandelt werden können. Die Anzahl der Kapitel ist deutlich geringer als die Anzahl von Lehrveranstaltungen in einem Semester. Dies liegt daran, dass sich die geschlossene Darstellung im Rahmen weniger Kapitel als didaktisch sinnvoller erwies als die einzelnen Themen zergliedert über mehrere Kapitel hinweg zu behandeln. Die einzelnen Kapitel weisen jedoch eine sehr strukturierte Binnengliederung auf, sodass insbesondere auch die umfangreicheren Kapitel 3, 4 und 6 in sinnvolle Lehreinheiten aufgeteilt werden können, um sie dem spezifischen Lehrplan anzupassen.

Am Entstehen des Buches haben viele mitgewirkt, denen wir ganz herzlich danken möchten. Angela Coenders hat viele Diktate zu Papier gebracht, Grafiken und Tabellen erstellt und Teile des Buches und insbesondere das Literaturverzeichnis korrigiert. Dr. Georg Hosoya hat nicht nur alle Kapitel des Buches korrekturgelesen und alle Beispiele zur Kontrolle nachgerechnet, sondern auch die kommentierten `R`-Inputs und -Outputs für die Internetseite des Buches erstellt. Claudia

Crayen hat R-Inputs für Grafiken erstellt. Sie hat darüber hinaus zusammen mit Dr. Luna Beck, Fenne große Deters, Nele Feuerbach, Frances Hoferichter, Louisa Hohmann, Dr. Tobias Koch, Dr. Irina Kumschick, Tanja Kutscher, Mario Lawes, Jana Mahlke, Martin Schultze und Linda Wulkau Teile des Manuskripts korrekturgelesen, auf Verständlichkeit geprüft sowie wesentliche Verbesserungsvorschläge vorgelegt. Linda Wulkau hat zentrale Teile der Formelsammlung erstellt. Ramzi Fatfouta, Henriette Hunold und Tanja Kutscher haben umfangreiche Literaturrecherchen durchgeführt und diese aufgearbeitet. Die beiden Mitherausgeber der Reihe, Prof. Dr. Franz Petermann und Prof. Dr. Hans-Werner Bierhoff, haben ebenfalls den Text sehr sorgfältig gelesen und wichtige Verbesserungshinweise gegeben. Ihnen allen möchten wir herzlich danken. Wir freuen uns auf weitere Hinweise und Verbesserungsvorschläge!

Berlin, im Januar 2014

Michael Eid
Katharina Schmidt

Kapitel 1

Grundfragen der Testtheorie und Testkonstruktion

Inhaltsübersicht

Aufgaben der psychologischen Diagnostik

Eine der wesentlichen Aufgaben der *psychologischen Diagnostik* besteht darin, die Merkmalsausprägungen von Menschen zu erfassen. Dies ist notwendig, um wichtige diagnostische Fragen beantworten und Entscheidungen treffen zu können (Eid & Petermann, 2006). Wenn z. B. ein Kind in einer schulpsychologischen Beratungsstelle vorgestellt wird, da es den Unterricht wiederholt stört, geht es im Wesentlichen zunächst darum, die Ursachen dieses Verhaltens festzustellen. Eine mögliche Ursache könnte eine unterdurchschnittliche Intelligenz sein, die zu einer Überforderung führt, oder aber eine überdurchschnittliche Intelligenz, aufgrund derer das Kind unterfordert ist und sich langweilt. In beiden Fällen wäre die Störung des Unterrichts ein Ausdruck der Bewältigung der emotionalen Belastung, die sich durch die Unter- bzw. Überforderung ergibt. Um geeignete Maßnahmen wie z. B. Nachhilfe oder die Empfehlung, eine Klasse zu überspringen, auswählen zu können, ist es entscheidend, die Ursache zu kennen. Deswegen würde die Schulpsychologin die Intelligenz des Kindes feststellen (diagnostizieren) wollen.

Da das Ergebnis der Intelligenzdiagnostik weitreichende Folgen für die Zukunft des Kindes haben kann, wird die Psychologin alles daran geben, das beste Verfahren zur Erfassung der Intelligenz auszuwählen. Doch wodurch zeichnet sich das beste Verfahren aus und wie kann die Psychologin in ihrem Gutachten belegen, dass sie auf das beste Verfahren zurückgegriffen hat und daher bei der Auswahl des Verfahrens keinen Fehler begangen hat? Hierzu muss sie auf Qualitätskriterien zurückgreifen, die ihr erlauben, aus der Vielzahl der verfügbaren Instrumente dasjenige auszuwählen, das für ihre Fragestellung am geeignetsten ist und die *Gütekriterien* erfüllt, die sich in der Wissenschaft und Praxis bewährt haben und allgemein anerkannt sind (Kersting, 2008; Kersting, Häcker & Hornke, 2011). So sollte dieses Instrument auch wirklich die Intelligenz erfassen und nicht ein anderes Merkmal wie z. B. die Testängstlichkeit. Ob ein Instrument wirklich das erfasst, was es erfassen soll, und ob Schlussfolgerungen, die wir aus den Ergebnissen in Bezug auf das zu erfassende Merkmal ziehen, auch wirklich zutreffend und damit gültig sind, betrifft die Frage der *Validität*, dem zentralen Gütekriterium der Diagnostik (Messick, 1989). Schlüsse, die aufgrund der Messungen mit einem Erhebungsinstrument getroffen werden, sollten daher über eine hohe Validität verfügen.

Gütekriterien

Validität

Das Erhebungsinstrument sollte darüber hinaus ermöglichen, die Intelligenz des Kindes zuverlässig zu erfassen. Wenn das Kind in kur-

zem Abstand mit dem Instrument erneut untersucht wird oder wenn ein anderes Instrument, das ebenfalls die Intelligenz valide misst, zur Feststellung der Intelligenz des Kindes herangezogen wird, sollten sich die Ergebnisse, z. B. der Intelligenzquotient des Kindes, zwischen den Messungen oder den Tests nur unwesentlich unterscheiden. Unterschiede zwischen den verschiedenen Messergebnissen zeigen an, dass die einzelne Messung mit einem Messfehler behaftet ist. Messfehler kann man bei psychologischen Erhebungen meist nicht ausschließen. Ihr Einfluss sollte aber möglichst gering sein. Ein Messinstrument, bei dem dies der Fall ist, ist ein zuverlässiges, man sagt auch reliables Messinstrument. Das dazugehörige Gütekriterium heißt *Reliabilität*. Das Instrument sollte zudem so geartet sein, dass das Ergebnis der Untersuchung – die diagnostizierte Intelligenz – nicht von der Schulpsychologin abhängt, die die Untersuchung durchgeführt hat. Eine andere Schulpsychologin sollte zu demselben Ergebnis kommen. Dieses Gütekriterium, das die Unabhängigkeit der Testergebnisse von der Person der Untersuchungsleiterin bzw. des Untersuchungsleiters fordert, nennt man *Objektivität*. Damit sind nur die wesentlichen Gütekriterien, die sogenannten Hauptgütekriterien, genannt. Man kann sich schon vorstellen, dass es nicht einfach und sehr aufwendig ist, ein objektives, reliables und valides Instrument zu entwickeln. Die praktisch tätige Schulpsychologin wird daher in den seltensten Fällen ein solches Instrument selbst erstellen, sondern aufgrund ihrer Expertise und ihrer wissenschaftlichen Ausbildung, die sie im Rahmen ihres Studiums in der Konstruktion und Bewertung psychologischer Erhebungsverfahren erhalten hat, ein adäquates Instrument auswählen. Ohne dieses theoretische Wissen wäre sie nicht in der Lage, ihren Beruf angemessen auszuüben.

Reliabilität

Objektivität

1.1 Was sind psychologische Messungen?

Wie konstruiert man nun ein psychologisches Erhebungsverfahren und wie kann man sicherstellen, dass es das erfasst, was es erfassen soll? Das vorliegende Buch vermittelt die wesentlichen theoretischen Grundlagen, die hierfür notwendig sind, und zeigt ihre praktischen Implikationen auf. Die Grundfragen, die sich hierbei stellen, sind ganz ähnlich zu den Fragen, die sich bei der Erfassung von Merkmalen auch in anderen wissenschaftlichen Disziplinen ergeben. So sind wir mit der Messung physikalischer Eigenschaften sehr vertraut. Die Messungen der Länge und des Gewichts sind uns geläufig und wir greifen fast täglich auf sie zurück. Sie sind uns so vertraut, dass wir

Vergleich zur physikalischen Messung

sie nicht hinterfragen und davon ausgehen, dass wir sie zuverlässig anwenden können. Aber auch sie basieren auf bestimmten Annahmen.

Größenmessung

Es gibt verschiedene Gemeinsamkeiten zwischen der Messung eines Merkmals in der Psychologie (z. B. Intelligenz) und der Messung eines Merkmals in der Physik. Dies kann am Beispiel der *Größenmessung* verdeutlicht werden: Wenn eine Ärztin die Größe eines ausgewachsenen Erwachsenen bestimmen will, benötigt sie einen Messapparat, der die Größe in der angegebenen Maßeinheit (z. B. Zentimeter) korrekt misst und nicht eine andere Eigenschaft wie z. B. das Gewicht (Validität). Der Messapparat muss sich dadurch auszeichnen, dass in sehr kurzen Zeitabständen wiederholte Größenmessungen desselben Erwachsenen zu ähnlichen – im Idealfall zu denselben – Ergebnissen führen (Reliabilität). Die wiederholten Größenmessungen werden wahrscheinlich nicht perfekt identisch sein, da der Erwachsene nicht immer gleich steht oder liegt, sodass allein dadurch eine gewisse Messungenauigkeit vorliegt. Ein Messapparat ist aber umso besser, je geringer diese Verfälschungen durch Messfehler sind. Die Messanordnung sollte daher so gestaltet sein, dass die Messung nicht durch die Person, die die Messung durchführt, systematisch verändert wird (Objektivität). Dies wäre z. B. dann nicht gegeben, wenn die Größenmessung im Stehen erfolgt und eine Arzthelferin, die die Messung durchführt, den Kopf eines Erwachsenen systematisch etwas nach unten drückt, eine andere Arzthelferin dies jedoch nicht tut.

Bei der Messung der Körpergröße lässt sich die Frage der Validität relativ einfach klären. Üblicherweise geschieht dies dadurch, dass die Ergebnisse der Größenmessung mit diesem Messinstrument mit einem Vergleichsstandard verglichen werden. Dies geschieht z. B. im Rahmen der Eichung eines Messinstruments. Unterscheiden sich die beiden Messinstrumente nur gering, spricht dies für die Validität des Größenmessinstruments. Die beiden anderen Gütekriterien (Reliabilität und Objektivität) sind an den Messvorgang, und zwar an das Messobjekt und die Personen, die die Messung durchführen, geknüpft.

Unterschiede zur physikalischen Messung

Unterschiede zwischen psychologischen und physikalischen Messungen. Die Erfassung von psychologischen Merkmalen verfolgt im Grunde dasselbe Ziel wie die Messung von Größen oder Längen, unterscheidet sich aber typischerweise von dieser in einigen wesentlichen Aspekten:

1. Bei der Erfassung psychologischer Merkmale gibt es keine allgemein akzeptierten Vergleichsstandards („golden standards"), die man zur Eichung von Messinstrumenten heranziehen könnte. Dies liegt u. a. daran, dass häufig eine allgemein akzeptierte Definition eines Merkmals fehlt und es auch keine normierten Maßeinheiten wie bei der Größen- oder Längenmessung gibt. Die international gebräuchliche Basiseinheit der Längenmessung ist der Meter, der definiert wird als die „Wegstrecke, die das Licht im Vakuum während einer Zeit von 1/299792458 Sekunde zurücklegt" (Giancoli, 2006, S. 8). In der Psychologie gibt es typischerweise keine vergleichbar präzisen Festlegungen einer Maßeinheit und des zu messenden Merkmals. Dies hat Konsequenzen für die Bestimmung der Validität. Ob ein Messinstrument das misst, was es messen soll, kann nicht anhand eines einfachen Vergleichs mit einem anderen Messinstrument bestimmt werden, sondern muss in meist sehr umfangreichen Studien untersucht werden. Da es sich bei der Frage nach der Validität um eine komplexe Forschungsfrage handelt, wurden verschiedene Validitätskonzepte entwickelt, zu deren Untersuchung verschiedene Forschungsstrategien geeignet sind.

Keine allgemein akzeptierten Vergleichsstandards für die Eichung von Messinstrumenten

2. In der Psychologie kann man das interessierende Merkmal anhand eines einzelnen Messvorgangs häufig nur sehr viel gröber messen, als dies bei der Größenmessung der Fall ist. So geht man bei der Intelligenz typischerweise davon aus, dass es sich bei dieser um ein kontinuierliches Merkmal handelt und somit sehr feine Unterschiede zwischen Personen in ihrer Intelligenz bestehen. Zur Messung der Intelligenz liegen aber häufig nur Aufgaben vor, die entweder korrekt gelöst oder nicht gelöst werden. Anhand einer einzelnen Aufgabe kann man Personen somit nur zwei Gruppen zuordnen, nämlich der Gruppe, die die Aufgabe gelöst, und der Gruppe, die die Aufgabe nicht gelöst hat. Will man die Intelligenz feiner bestimmen, müssen mehrere Aufgaben bearbeitet werden. Dann stellt sich aber zwangsläufig die Frage, ob die verschiedenen Aufgaben dasselbe Merkmal (Intelligenz) messen oder möglicherweise verschiedene Merkmale. Die Fähigkeit, Zahlen zu multiplizieren, kann über Aufgaben wie

Mit welchen Aufgaben soll das psychologische Merkmal gemessen werden?

Beispiel Intelligenzmessung

$$2 \cdot 9 = ?$$

und

Franz hat zwei Spielzeugautos und sein
Bruder Fritz hat neunmal so viele Spielzeugautos.
Wie viele Spielzeugautos hat Fritz?

erfasst werden. Obwohl beide Aufgaben auf dieselbe rechnerische Fähigkeit abzielen, kann nicht ungeprüft davon ausgegangen werden, dass sie auch wirklich dasselbe Merkmal erfassen. Die korrekte Lösung der zweiten Aufgabe hängt möglicherweise nicht nur von der Multiplikationsfähigkeit, sondern auch vom Textverständnis ab. Dann würden beide Aufgaben partiell unterschiedliche Merkmale erfassen. Dies müsste bei der Diagnostik der rechnerischen Fähigkeit berücksichtigt werden. Beide Aufgaben wären dann nicht austauschbar und könnten nicht einfach zur Messung desselben Merkmals herangezogen werden. Die Frage, ob verschiedene Aufgaben dasselbe Merkmal erfassen, ist also eine äußerst wichtige Frage der Konstruktion einer Erfassungsmethode.

Messen verschiedene Aufgaben dasselbe Merkmal?

Genauso relevant ist aber auch die Frage, wie viele Aufgaben benötigt werden und wie die Aufgaben im Einzelnen gestaltet werden sollen, um eine zuverlässige Bestimmung der Merkmalsausprägung einer Person zu ermöglichen. Es liegt nahe zu vermuten, dass die Erfassung umso präziser ist, je mehr Aufgaben man vorgibt. Aber ist dies immer so? Wovon hängt die Präzision ab? Ist es günstiger, Aufgaben vorzugeben, die ähnlich sind oder sich stark unterscheiden? Wie kann man sicherstellen, dass man eine gewünschte Präzision auch wirklich erhält, und wie kann man Aufgaben so zusammenstellen, dass Personen einerseits zeitlich nicht zu sehr beansprucht werden, ihre Merkmalsausprägungen andererseits aber hinreichend präzise erfasst werden? Um Antworten auf diese relevanten Fragen geben zu können, wurden mathematische Modelle entwickelt, die der Konstruktion und Analyse von Erfassungsmethoden in der Psychologie zugrunde gelegt werden können. Die Teildisziplin der Psychologie, die sich mit der Messung psychologischer Merkmale beschäftigt, ist die *Psychometrie*. Psychometrische Modelle erlauben es zu überprüfen, ob verschiedene Aufgaben dasselbe Merkmal messen und mit welcher Präzision sie dies tun.

Präzision

Psychometrie

Wir wollen zunächst die Grundidee eines solchen psychometrischen Modells darstellen, um aufzuzeigen, welche weiteren theoretischen und praktischen Fragen sich hieraus ergeben. Wir greifen auf das Rasch-Modell zurück, eines der einfachsten psychometrischen Modelle, das wir in Kapitel 4 im Detail behandeln werden. Wir werden dieses Modell anhand einer inhaltlichen Anwendung, der Erfassung der emotionalen Kompetenz, illustrieren.

1.2 Grundidee psychometrischer Modelle am Beispiel des Rasch-Modells

Eine wichtige Facette der emotionalen Kompetenz ist die Fähigkeit, den Emotionsausdruck von Personen korrekt einzuschätzen (Saarni, 2002). Zur Erfassung dieser Kompetenz werden Personen z. B. Bilder vorgelegt, auf denen das Gesicht eines Menschen abgebildet ist, der gerade eine Emotion erlebt. Die Personen sollen angeben, welche Emotion dieser Mensch gerade erlebt. Grob, Meyer und Hagmann-von Arx (2009) haben ein solches Verfahren für Kinder entwickelt. Ihre *Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren* (IDS) enthalten eine Teilskala mit dem Namen *Emotionen erkennen*. Dem Kind werden Bilder von zehn Personen vorgelegt, die eines von fünf verschiedenen Gefühlen zeigen. Wird die richtige Emotion genannt, erhält das Kind einen Punkt, wird die falsche Emotion angegeben, erhält das Kind keinen Punkt. Die Bilder für die Emotion Freude sind in Abbildung 1.1 dargestellt (Items 1 und 10). Hier wird beispielsweise ein Punkt für die Antworten „Freude haben", „froh", „lachen" etc. vergeben, und null Punkte für „freundlich", „gut", „schön" etc. Im Test werden Bilder zu folgenden weiteren Emotionen vorgelegt: Wut (Nr. 2 und Nr. 8), Angst (Nr. 3 und Nr. 7), Trauer (Nr. 4 und Nr. 6) und Überraschung (Nr. 5 und Nr. 9).

Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren

Abbildung 1.1: Items für die Emotion *Freude* in der Subskala *Emotionen erkennen* der *Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren* (Grob et al., 2009; Abdruck mit freundlicher Genehmigung des Verlags Hans Huber).

Um die Kompetenz eines Kindes, Emotionen in Gesichtern zu erkennen, zu bestimmen, wird die Anzahl der gelösten Aufgaben (Summenwert) ausgezählt. Ein Kind kann somit minimal 0 Punkte und maximal 10 Punkte erhalten. Je höher der Wert, umso höher wird die Kompetenz, Emotionen zu erkennen, eingeschätzt. Um zu bewerten, ob dieser Summenwert ein sinnvolles Maß der Messung der emotionalen Kompetenz ist, müssen zwei Fragen geklärt werden:

Summenwert als Kompetenzmaß

1. *Sprechen die verschiedenen Bilder dieselbe emotionale Kompetenz an?* Möglicherweise gibt es verschiedene emotionale Kompetenzen. So könnte sich z. B. die Fähigkeit, positiv bewertete Emotionen (z. B. Freude) zu erkennen, von der Fähigkeit, negativ bewertete Emotionen (z. B. Wut, Angst, Trauer) zu erkennen, unterscheiden. Wenn unterschiedliche Bilder unterschiedliche Fähigkeiten ansprechen, dann wäre die Bildung eines solchen Summenwertes problematisch, da unterschiedliche Fähigkeiten undifferenziert zusammen verrechnet würden. Man könnte dann z. B. an einem Wert von 5 nicht erkennen, ob diese mittlere Kompetenzausprägung nur darauf zurückgeführt werden kann, dass vor allem eine Gruppe von Emotionen (z. B. negative) nicht richtig erkannt werden, die anderen jedoch nahezu perfekt. Ein Wert von 5 könnte aber auch bedeuten, dass die Erkennungsfähigkeit bei allen Typen von Emotionen mittelmäßig ist.
2. *Sollen alle Aufgaben gleich gewichtet werden?* Berechnet man die Summe der gelösten Aufgaben, fließt jede Aufgabe mit gleichem Gewicht ein. Man könnte aber argumentieren, dass die Lösung einer schwierigeren Aufgabe ein stärkeres Gewicht bei der Bestimmung der Kompetenz bekommen sollte als die Lösung einer leichten Aufgabe. Möglicherweise gibt es Emotionen, die leichter, und solche, die schwieriger im Gesicht zu erkennen sind. Sollten in diesem Falle die Aufgaben unterschiedlich gewichtet werden und wenn ja, in welcher Weise?

Grundzüge des Rasch-Modells

Zur Klärung beider Fragen kann die Analyse von Daten mit psychometrischen Modellen einen wichtigen Beitrag leisten. Um beiden Fragen nachzugehen, wurden 193 Kindern die zehn Bilder vorgelegt, ihre Antworten bewertet und die so erhaltenen Daten mit dem Rasch-Modell analysiert. Das Rasch-Modell wurde von dem dänischen Statistiker Georg Rasch entwickelt und im Jahre 1960 in dem Buch *Probabilistic models for some intelligence and attainment tests* publiziert. Die Ergebnisse der Analyse mit dem Rasch-Modell sind in Abbildung 1.2 für die Aufgaben zum Erkennen von Emotionen in Gesichtern aus den IDS (Grob et al., 2009; siehe oben) dargestellt. In diesem Beispiel haben wir für die grafische Darstellung nur die ersten fünf Items ausgewählt, die jeweils eine andere Emotion zeigen. Was bedeutet diese Abbildung? Die Abszisse (x-Achse) ist mit *Fähigkeit, Emotionen zu erkennen* bezeichnet. Die Abszisse stellt das Merk-

Rasch-Modell

mal dar, das wir mit unseren Aufgaben erfassen wollen. Im Rasch-Modell ist dies eine kontinuierliche Variable, die Werte zwischen $-\infty$ und ∞ annehmen kann. Man geht somit davon aus, dass es sehr feine Unterschiede zwischen Personen in dieser emotionalen Kompetenz gibt. Die Ordinate (y-Achse) ist mit *Lösungswahrscheinlichkeit* bezeichnet und ist auf den Wertebereich von Wahrscheinlichkeiten (0 bis 1) begrenzt.

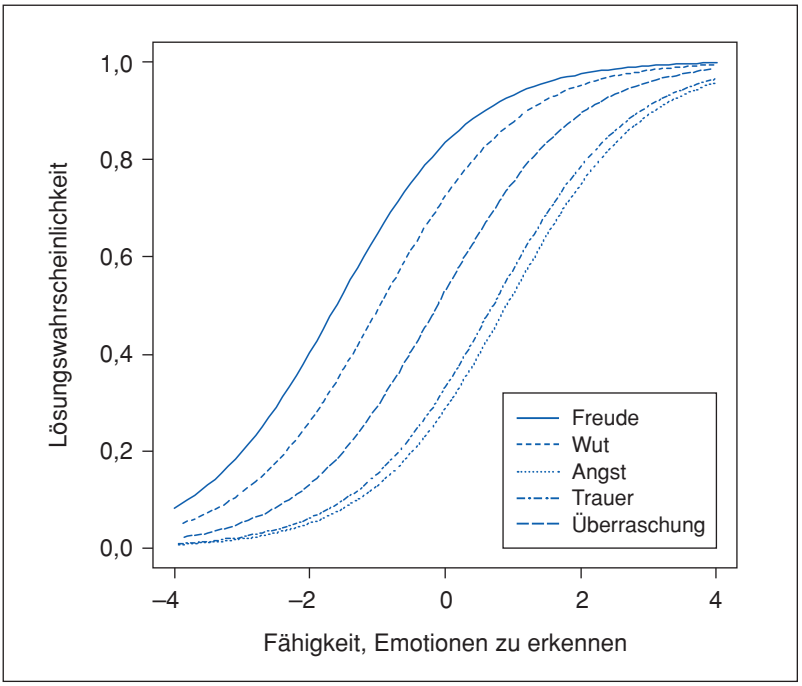

Abbildung 1.2: Ergebnis einer Analyse der Subskala *Emotionen erkennen* der *Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren* (Grob et al., 2009) mit dem Rasch-Modell. Dargestellt sind die Itemcharakteristiken der ersten fünf Aufgaben.

Die Grafik enthält weiterhin fünf s-förmige Kurven, eine für jede Aufgabe. Die Kurven repräsentieren die Lösungswahrscheinlichkeiten für jede Aufgabe in Abhängigkeit von der Fähigkeit, Emotionen zu erkennen. Diese Kurven werden *Itemcharakteristiken* genannt, da sie die Aufgabe (das *Item*) kennzeichnen (charakterisieren). Man kann leicht erkennen, dass die Lösungswahrscheinlichkeit mit Zunahme der emotionalen Kompetenz anwächst. Je höher die emotionale Kompetenz, desto größer ist die Wahrscheinlichkeit, eine Emotion korrekt

Itemcharakteristik

zu erkennen. Das Rasch-Modell geht somit davon aus, dass man die Lösung einer Aufgabe aufgrund der emotionalen Kompetenz nicht perfekt vorhersagen kann. Die Vorhersage ist nur mit einer bestimmten Wahrscheinlichkeit möglich und daher mit einer Unsicherheit verknüpft. Die Itemcharakteristiken unterscheiden sich in ihrer Lage auf der Abszisse, sie weisen aber alle die gleiche Form auf. Ganz links liegt die Kurve des ersten Bildes (Freude). Dieses Item ist leichter als alle anderen Items, da seine Itemcharakteristik für alle Werte der emotionalen Kompetenz größere Werte annimmt als die Itemcharakteristiken der anderen Items. Dies bedeutet, dass alle Personen – unabhängig von ihrer emotionalen Kompetenz – eine höhere Wahrscheinlichkeit aufweisen, die Emotion im ersten Bild (Freude) korrekt zu erkennen, als in den anderen Bildern. Am schwierigsten ist das dritte Item (Angst). Die Wahrscheinlichkeit, die Emotion in diesem Bild korrekt zu erkennen, ist für alle Personen geringer als bei den anderen Items. Die Itemcharakteristiken der anderen Items liegen dazwischen. Die Wahrscheinlichkeit, eine Emotion korrekt zu erkennen, hängt also sowohl von der Fähigkeit einer Person als auch der Schwierigkeit einer Aufgabe ab. Je größer die Fähigkeit einer Person und je geringer die Schwierigkeit einer Aufgabe, umso größer ist die Lösungswahrscheinlichkeit.

Personenfähigkeit und Itemschwierigkeit

Wie der Name schon sagt, handelt es sich bei dem Rasch-Modell um ein *Modell*, also eine hypothetische Vorstellung darüber, wie die Lösungswahrscheinlichkeiten verschiedener Aufgaben von einem Merkmal abhängen. Es geht davon aus, dass die Lösungswahrscheinlichkeiten aller Items (neben der Itemschwierigkeit) von nur einem Personenmerkmal (Fähigkeit, Emotionen zu erkennen) abhängen und dass alle Items dasselbe Personenmerkmal messen. Dieses Personenmerkmal ist nicht direkt beobachtbar, sondern liegt dem Verhalten zugrunde. Es wird über die beobachtbaren Verhaltensweisen erschlossen.

Annahmen und Eigenschaften psychometrischer Modelle

Begriffsklärungen

Latente Variable

Latentes Merkmal oder latente Variable: Ein Personenmerkmal, dessen Ausprägungen nicht direkt beobachtbar sind, sondern über beobachtbares Verhalten erschlossen werden.

Manifeste Variable

Beobachtbare oder manifeste Variable: Im Gegensatz zu latenten Variablen sind die Ausprägungen einer manifesten Variablen direkt beobachtbare Reaktionen (z. B. Lösen vs. Nichtlösen einer Aufgabe).