Wissen Frauen weniger oder nur das Falsche? Ein statistisches Modell für unterschiedliche Aufgaben-Schwierigkeiten in Teilstichproben

Carolin Strobl, Julia Kopf (Ludwig-Maximilians-Universität München), Achim Zeileis (Leopold-Franzens-Universität Innsbruck)

Zusammenfassung:

Eines der erstaunlichsten Ergebnisse des Studentenpisa-Tests ist das deutlich schlechtere Abschneiden der Frauen. Aber haben Frauen wirklich ein schlechteres Allgemeinwissen, oder hat der Test ihnen nur keine faire Chance eingeräumt? Unfair wäre der Test, wenn Männer kein besseres Allgemeinwissen hätten, aber trotzdem im Test besser abschneiden würden, weil ihnen bestimmte Aufgaben leichter fallen. Diese Möglichkeit steht - wie bereits im vorherigen Beitrag - im Zentrum der Fragestellung, wird hier allerdings mit einem neuen statistischen Verfahren überprüft. Dieses Verfahren basiert auf einer Kombination des Rasch-Modells, das auch bei der Auswertung der offiziellen PISA-Studie eingesetzt wird, mit Methoden aus der Wirtschaftsstatistik. Im Ergebnis zeigt sich, dass der Studentenpisa-Test durchaus Fragen enthält, die Männer bei gleichem Allgemeinwissen leichter fallen als Frauen, aber dieser Effekt allein das schlechtere Abschneiden der Frauen nicht erklären kann.

1. Gruppenunterschiede in Leistungstests – eine hitzige Debatte

Das unterschiedliche Abschneiden von verschiedenen Personengruppen – insbesondere von Frauen und Männern – in Intelligenz- und Leistungstests führt seit der Entwicklung der ersten Intelligenztests im frühen 20. Jahrhundert zu hitzigen Debatten unter Wissenschaftlern (siehe z.B. die aktuelle Gegenüberstellung von Asendorpf und Wenderlein 2009). Inhaltlich erschwert wird diese Debatte dadurch, dass es sich bei eventuellen Unterschieden nicht notwendigerweise um „echte" Leistungsunterschiede handelt. Vielmehr gibt es eine Vielzahl von Einflussgrößen, wie z.B. Sozialisation und Bildungshintergrund, unterschiedliche Lösungsstrategien und Vorerfahrungen mit ähnlichen Tests, aber auch die Konstruktion der Tests selbst, die sich auf die erreichte Punktezahl auswirken können.

Mögliche Quellen von Gruppenunterschieden sind unterschiedliche Motivation (Simpson und Oliver 1985) und Strategien bei der Bearbeitung des Tests (Ben-Shakhar und Sinai 1991). So könnte es z.B. sein, dass Männer von sich aus eine höhere Motivation haben, in einem Wissenstest gut abzuschneiden und sich entsprechend mehr anstrengen, während Frauen nur „zum Spaß" teilnehmen. Ein methodisch schwieriger Punkt ist die unterschiedliche Vorbildung von Personengruppen in unterschiedlichen Themenbereichen, die hier zunächst alle unter dem gemeinsamen Konstrukt „Allgemeinbildung" gefasst werden sollen – insbesondere auch das unterschiedliche Interesse für verschiedene Themengebiete, aufgrund dessen Wissen als mehr oder weniger relevant eingestuft, und entsprechend vielleicht mehr oder weniger erfolgreich aufgenommen, behalten und wiedergegeben wird (Woike, Bender und Besner 2008).

Wir werden später genauer untersuchen, ob die Geschlechterunterschiede im Studentenpisa in bestimmten Themenbereichen besonders groß sind. Eine ähnliche Problematik ergibt

sich aber auch bei Intelligenz- und Schulleistungstests mit unterschiedlichen Themenbereichen: Setzt sich ein Test z.B. aus verbalen und mathematischen Aufgaben zusammen, zeigt sich oft ein Leistungs-(und auch Motivations-)vorsprung für Frauen bzw. Mädchen in der Leseleistung (Grütz 2004) und für Männer bzw. Jungen in den mathematischen Aufgaben. Die Ursachen für diese Unterschiede können in der Sozialisation und geschlechtsspezifischen Rollenerwartungen liegen: Mädchen haben ein geringeres mathematisches Selbstvertrauen und ihr Interesse für mathematische und naturwissenschaftliche Fächer wird vom sozialen Umfeld weniger gefördert (Jahnke-Klein 2005; Schnurr 2007). Umgekehrt hat sich in den letzten Jahren eine angeregte Diskussion darüber entwickelt, ob Jungen im Schulsystem benachteiligt und sprachlich zu wenig gefördert werden (vgl. z.B. Trenkamp 2009).

Bei den meisten Intelligenz- und Schulleistungstests verschwinden die Geschlechterunterschiede, wenn man z.B. aus den verbalen und mathematischen Aufgaben einen gemeinsamen Score berechnet, da sich Stärken und Schwächen beider Gruppen bei diesen beiden Themengebieten ausgleichen. Allerdings weisen die Scores von Männern oft eine breitere Streuung auf, so dass sich unter den besten Probanden besonders viele Männer finden (Hedges and Nowell 1995) – die breitere Streuung kann aber (bei gleichem Mittelwert und symmetrischer Verteilung) gleichzeitig dazu führen, dass auch besonders viele Männer unterdurchschnittlich abschneiden.

Eine weitere Problematik ergibt sich dadurch, dass es sehr schwer ist, einen Test zu konstruieren, der nur das gewünschte Merkmal (aber nichts anderes) misst. Auch bei einem Test zu mathematischen Fähigkeiten werden die Aufgaben z.B. als Textaufgaben formuliert. Entsprechend kann es vorkommen, dass bestimmte Aufgaben z.B. für Probanden mit schlechteren Sprachkenntnissen schwieriger zu lösen sind. In all solchen Fällen spricht man von „Differential Item Functioning" (Differentielle Itemfunktion, im Folgenden mit DIF abgekürzt): Die Schwierigkeit einer Aufgabe unterscheidet sich für bestimmte Gruppen von Probanden – und zwar nicht aufgrund von tatsächlich unterschiedlichen (z.B. mathematischen) Fähigkeiten, sondern aus anderen Gründen. Dadurch kann auch DIF dazu führen, dass eine Gruppe von Personen in einem Test schlechter abschneidet. Diese mögliche Ursache für Gruppenunterschiede soll im Folgenden für das Studentenpisa weiter untersucht werden.

2. Das Rasch-Modell zur Auswertung psychologischer Tests

Eines der wichtigsten Modelle zur Auswertung psychologischer Tests ist das Rasch-Modell, das von dem dänischen Statistiker Georg Rasch entwickelt wurde (Rasch 1960). Das Rasch-Modell wird u.a. bei der Auswertung der „echten" PISA-Studie verwendet, um die Leistungen von Schülern aus unterschiedlichen Ländern zu messen.

Das Rasch-Modell ist ein mathematisches Modell, das Parameter für die Fähigkeiten der getesteten Personen, aber auch für die Schwierigkeiten der verwendeten Aufgaben enthält. Wenn ein Test den strengen mathematischen Anforderungen des Rasch-Modells genügt, sind dadurch objektive Vergleiche zwischen Personen, aber auch zwischen Aufgaben, möglich. Ein psychologischer Test, der den mathematischen Anforderungen des Rasch-Modells genügt, muss u.a. folgende Bedingungen erfüllen: Die Antworten der Personen müssen unabhängig voneinander gegeben werden. Diese Annahme wäre z.B. verletzt, wenn Personen voneinander abschreiben. Auch die Aufgaben müssen unabhängig voneinander lösbar sein. Diese Annahme wäre z.B. verletzt, wenn Aufgaben so aufeinander aufbauen, dass für die Lösung einer Auf-

gabe das Ergebnis einer anderen Aufgabe nötig ist. Außerdem darf die Lösung der Aufgabe nur von einer einzigen Fähigkeit abhängen.

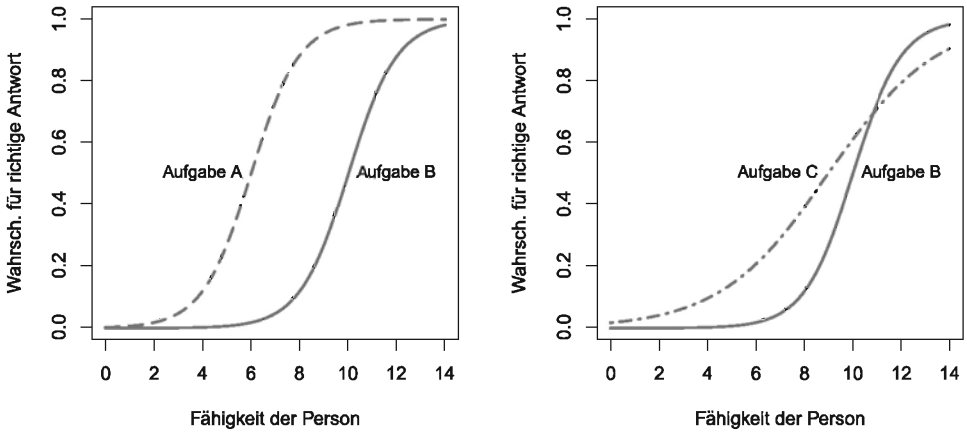
Dieser letzte Punkt bedeutet zum einen, dass das Rasch-Modell in seiner ursprünglichen Form nur ein eindimensionales latentes (d.h. nicht direkt beobachtbares) Merkmal messen kann. Bei Tests zur Intelligenz- und Leistungsmessung, aber auch bei Wissenstests, lässt sich das übergeordnete Konstrukt hingegen oft in verschiedene Teilbereiche wie z.B. mathematische und verbale Fähigkeiten unterteilen. Dieser Unterteilung liegt sowohl eine theoretische Trennung der Teilbereiche zugrunde, als auch die empirische Beobachtung, dass es Personen gibt, die in einem der Bereiche gut, und im anderen hingegen schlecht abschneiden. In einem Intelligenz- oder Leistungstest könnte es also zwei Teilbereiche mit mathematischen und verbalen Aufgaben geben, die man gerne komplett voneinander trennen würde. In der Realität hängt die Beantwortung einer Aufgabe aber oft von mehreren Faktoren ab. Man unterscheidet dabei zwischen dem oben beschriebenen DIF, bei dem ein Merkmal gemessen werden soll, aber andere unerwünschte Einflüsse sich auf die Messung auswirken, und mehrdimensionalen Modellen, bei denen tatsächlich mehrere Merkmale gemessen werden sollen.

Hier folgen wir aber zunächst der ursprünglichen Auswertung im SPIEGEL-Artikel zum Studentenpisa (Verbeet 2009), wo für jede Person die Anzahl der gelösten Aufgaben über alle Themengebiete als Indikator für Allgemeinbildung betrachtet wurde. Diese Herangehensweise impliziert ein eindimensionales Rasch-Modell über alle Themengebiete.

Wir gehen also davon aus, dass alle Aufgaben ausschließlich dieselbe latente Eigenschaft „Allgemeinbildung" messen. Das Rasch-Modell besagt dann, dass die Wahrscheinlichkeit, dass eine Person eine Aufgabe richtig löst, von der Fähigkeit (d.h. hier der Allgemeinbildung) der Person und der Schwierigkeit der Aufgabe abhängt: Je höher die Fähigkeit der Person, desto höher die Wahrscheinlichkeit, eine Aufgabe richtig zu lösen. Für alle Personen haben aber natürlich schwierigere Aufgaben eine niedrigere Lösungswahrscheinlichkeit als leichtere Aufgaben.

Die Lösungswahrscheinlichkeiten von zwei Aufgaben sind in Abbildung 1 (links) dargestellt. Die leichtere Aufgabe A ist für alle Personen mit höherer Wahrscheinlichkeit zu lösen als die schwierigere Aufgabe B. Das erkennt man daran, dass die Kurve für Aufgabe A immer oberhalb der Kurve für Aufgabe B verläuft.

Abbildung 1: Lösungswahrscheinlichkeiten von Aufgaben, die zum Rasch-Modell passen



(links): Die Lösungswahrscheinlichkeit für die schwierigere Aufgabe B ist immer niedriger als die Lösungswahrscheinlichkeit für die leichtere Aufgabe A. Im Vergleich dazu Lösungswahrscheinlichkeiten von Aufgaben, die nicht zum Rasch-Modell passen (rechts): Aufgabe C wird von Personen mit geringer Fähigkeit mit höherer Wahrscheinlichkeit gelöst als Aufgabe B; von Personen mit hoher Fähigkeit wird hingegen Aufgabe B mit höherer Wahrscheinlichkeit gelöst. Man kann also nicht mehr generell sagen, welche Aufgabe leichter ist.

Eine Verletzung des Rasch-Modells könnte man hingegen daran erkennen, dass sich die Kurven für zwei Aufgaben schneiden, wie in Abbildung 1 (rechts). Inhaltlich würde dies bedeuten, dass Aufgabe C für Personen mit niedrigerer Fähigkeit mit höherer Wahrscheinlichkeit zu lösen ist als Aufgabe B. Für Personen mit höherer Fähigkeit ist Aufgabe C hingegen mit niedrigerer Wahrscheinlichkeit zu lösen als Aufgabe B. Auch für jede andere Einteilung der Personen, z.B. nach Geschlecht oder Bildungsstand, muss gelten: Dieselbe Aufgabe muss für alle Personen entweder leichter oder schwieriger als eine Vergleichsaufgabe sein – sie darf nicht für manche Personen leichter und für andere schwieriger sein. Diese zentrale Annahme des Rasch-Modells bezieht sich also nicht auf die absolute Schwierigkeit einer Aufgabe, sondern auf den Vergleich der Schwierigkeiten von mehreren Aufgaben, der für alle Personen gleich ausgehen muss.

Diese Anforderung kann bei einem echten psychologischen Test natürlich durchaus verletzt sein. Ist dies der Fall, sind mit diesem Test objektive und damit faire Vergleiche zwischen unterschiedlichen Gruppen wie im Rasch-Modell nicht möglich. Insofern ist es kein Nachteil sondern ein Vorteil des Rasch-Modells, dass es so strenge Forderungen stellt: Die Annahmen können mithilfe statistischer Tests überprüft werden, und damit kann eine objektive Messung sichergestellt werden – wohingegen man ohne das Modell gar nicht merken würde, wenn die Messung nicht objektiv ist.

3. Anwendung des Rasch-Modells beim Studentenpisa

Die Probanden des Studentenpisa bekamen zufällig einen Fragebogen mit 45 Aufgaben aus den fünf Themengebieten Politik, Geschichte, Wirtschaft, Kultur und Naturwissenschaften zugelost. Bereits bei der Auswertung der Rohwerte (d.h. der Anzahl richtiger gelöster Aufgaben) im SPIEGEL-Artikel (Verbeet 2009) fielen dabei Aufgaben auf, die z.B. selektiv von Männern oder Frauen bzw. von älteren Probanden besonders häufig richtig gelöst wurden. Im Folgenden werden wir beispielhaft einen einzelnen Fragebogen mit 45 Aufgaben betrachten, der von 30188 Personen beantwortet wurde. Der Fragebogen wurde danach ausgewählt, dass er die meisten (insgesamt vier) der im SPIEGEL-Artikel genannten Aufgaben enthält, die für einzelne Personengruppen leichter zu lösen waren (z.B. die Aufgabe „Wie heißt der Bestsellerroman von Daniel Kehlmann?" – ‚Die Vermessung der Welt', die für Frauen leichter zu lösen war als für Männer, sowie die Aufgabe „Wer ist das?" – Foto von Daimler-Vorstand Dieter Zetsche', die für Männer leichter zu lösen war als für Frauen). Im Folgenden werden wir mithilfe des Rasch-Modells formal nachweisen, dass die im Artikel genannten und viele weitere Aufgaben signifikantes DIF aufweisen.

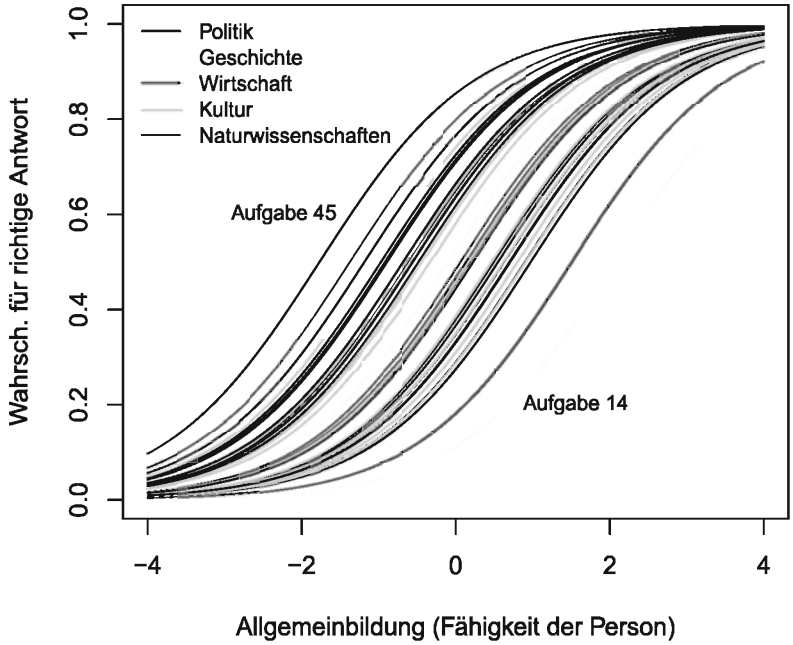
Im ersten Schritt der statistischen Analyse wird dazu für alle Personen ein gemeinsames Rasch-Modell angepasst. Das Modell ergibt für jede Person eine Schätzung ihres Allgemeinwissens. Dadurch lassen sich die Beobachtungen des SPIEGEL-Artikels replizieren, aber auch die Annahmen des Rasch-Modells, die für objektive Vergleiche zwischen Personengruppen nötig sind, überprüfen. In einem zweiten Schritt werden wir dann untersuchen, welche Aufgaben DIF aufweisen, und ob dies als Ursache für die gefundenen Unterschiede in der Allgemeinbildung in Frage kommt. Neben der gezielten Überprüfung der Unterschiede zwischen Frauen und Männern kommt dabei auch ein neues statistisches Verfahren zum Einsatz, das Gruppen von Personen mit unterschiedlichen Aufgaben-Schwierigkeiten automatisch aufspürt.

3.1. Ein gemeinsames Rasch-Modell

Abbildung 2 zeigt die geschätzten Lösungswahrscheinlichkeiten für die 45 Aufgaben des Fragebogens, wenn man das Rasch-Modell zugrunde legt und alle Probanden gemeinsam untersucht. Die über alle Personen hinweg betrachtet schwierigste Aufgabe in diesem Fragebogen stammt aus dem Themenbereich Geschichte („Wie viel Prozent der Wählerstimmen erhielt die NSDAP bei der Reichstagswahl im Jahr 1928?" – rund drei Prozent). Ähnlich schwierig sind zwei Aufgaben aus den Bereichen Politik („Wer bestimmt in Deutschland laut Grundgesetz die ‚Richtlinien der Politik' ?" – der Bundeskanzler bzw. die Bundeskanzlerin) und Wirtschaft („Wer ist das?" – Foto von Daimler-Vorstand Dieter Zetsche). Die leichteste Aufgabe in diesem Fragebogen ist aus dem Themenbereich Naturwissenschaften („Was ist die Summe der Innenwinkel eines Dreiecks?" – 180 Grad).

Zu beachten ist, dass die Form der Kurven in Abbildung 2 noch nicht bedeutet, dass die Annahmen des Rasch-Modells erfüllt sind. Sie stellen nur die unterschiedlichen Schwierigkeiten der Aufgaben dar, die sich aus einem Rasch-Modell ergeben würden. Ob die für das Rasch-Modell nötigen Annahmen für das Studentenpisa wirklich gerechtfertigt sind, untersuchen wir später noch genauer.

Abbildung 2: Lösungswahrscheinlichkeiten für die 45 Aufgaben des Fragebogens aus den Themenbereichen Politik, Geschichte, Wirtschaft, Kultur und Naturwissenschaften.

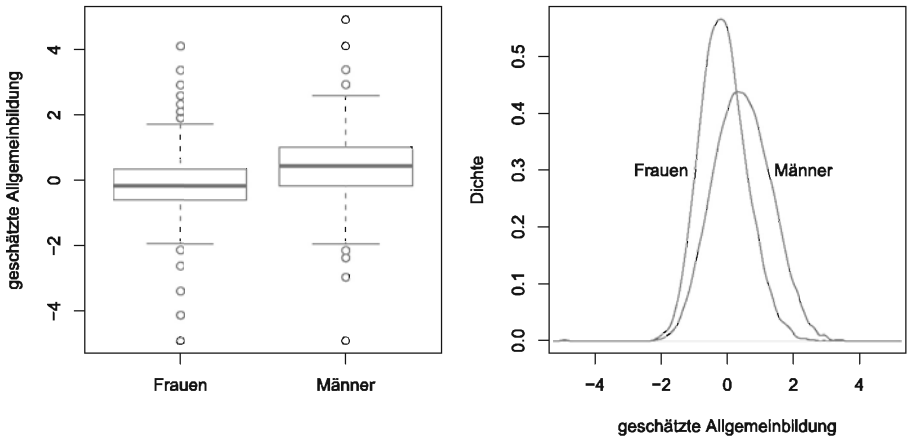


Die schwierigste Aufgabe (mit der niedrigsten Lösungswahrscheinlichkeit) Nr. 14 ist: „Wie viel Prozent der Wählerstimmen erhielt die NSDAP bei der Reichstagswahl im Jahr 1928?" aus dem Themenbereich Geschichte. Die leichteste Aufgabe (mit der höchsten Lösungswahrscheinlichkeit) Nr. 45 ist: „Was ist die Summe der Innenwinkel eines Dreiecks?" aus dem Themenbereich Naturwissenschaften.

3.2. *Unterschiede in der nach diesem Modell berechneten Allgemeinbildung*

Wie sich bereits im SPIEGEL-Artikel anhand der Rohwerte abgezeichnet hat, gibt es Unterschiede zwischen der nach einem gemeinsamen Rasch-Modell geschätzten Allgemeinbildung von Frauen und Männern: In Abbildung 3 ist die Verteilung der geschätzten Fähigkeiten für Frauen und Männer dargestellt. Die linke Darstellung enthält „Boxplots" – eine grafische Darstellung, die besonders gut die Lage einer Verteilung veranschaulicht: Der dicke Strich in der Mitte der „Box" gibt den Median der jeweiligen Gruppe an (d.h. den Wert, den mindestens die Hälfte der Personen in der Gruppe erzielt haben). Der Median ist – wie der Mittelwert – ein Maß für die Lage der Werte, hat aber gegenüber dem Mittelwert den Vorteil, dass einzelne besonders kleine oder große Werte, so genannte Ausreißer, ihn weniger stark beeinflussen. Die Länge der „Zäune", die im Boxplot nach oben und unten abgehen, sind hingegen ein Indikator für die Streuung der Werte (d.h. wie stark sich die Werte der einzelnen Personen unterscheiden). In Abbildung 3 sieht man, dass der Median in der Gruppe der Männer höher liegt, d.h. Männer haben im Mittel höhere geschätzte Fähigkeiten.

Abbildung 3: Unterschiede in der nach dem gemeinsamen Modell berechneten Allgemeinbildung von Frauen und Männern.



Die Streuungen der beiden Gruppen sind dagegen mithilfe der Boxplots nur schwer zu vergleichen. Dazu bietet sich die rechte Darstellung der geschätzten Dichtefunktion an. In dieser Darstellung erkennt man, dass die Verteilung für die Männer nicht nur nach rechts verschoben sondern auch breiter ist als die der Frauen, d.h. die Werte der Männer unterscheiden sich untereinander stärker als die Werte der Frauen – ein Ergebnis, das sich mit den Befunden von Hedges und Nowell (1995) zur größeren Streuung in den Scores von männlichen Probanden bei Leistungstests deckt.

3.3. *Überprüfung der Modellannahmen*

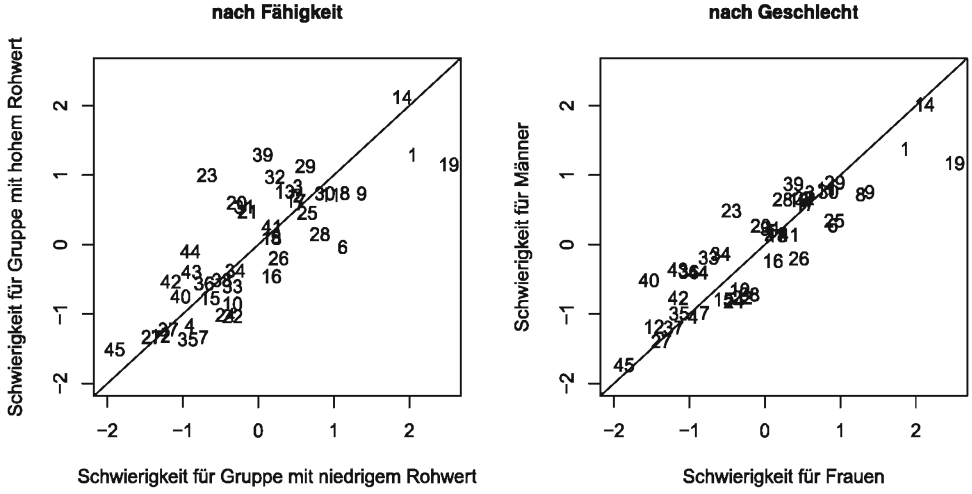
Die einfachste Methode zur Überprüfung der zentralen Modellannahme, dass dieselbe Aufgabe im Vergleich zu den übrigen Aufgaben nicht für manche Personen einfacher zu lösen sein darf als für andere, ist der grafische Modelltest. Dabei werden die Personen in zwei Gruppen aufgeteilt – nach ihrem Rohwert, d.h. der Anzahl der richtig gelösten Aufgaben, oder nach einem anderen Merkmal wie z.B. dem Geschlecht – und die geschätzten Aufgaben-Schwierigkeiten in beiden Gruppen verglichen: Wie in Abbildung 4 dargestellt, werden die geschätzten Aufgaben-Schwierigkeiten der zwei Gruppen in Richtung der x- und y-Achse abgetragen. Stimmen die Schätzungen überein, liegen alle Punkte auf der Winkelhalbierenden. Liegen die Punkte für einzelne Aufgaben oberhalb der Winkelhalbierenden, ist die Aufgabe für die nach oben abgetragene Gruppe schwieriger; liegen sie unterhalb Winkelhalbierenden, ist die Aufgabe für die nach oben abgetragene Gruppe leichter.

Beide Abbildungen zeigen Aufgaben, die von der Winkelhalbierenden abweichen, die also DIF zeigen. Auch der sog. Likelihood-Quotienten-Modelltest (ein gebräuchlicher statistischer Modelltest, der auch auf Unterschiede in den Schwierigkeits-Parametern von vorgegebenen Gruppen reagiert) spricht in beiden Fällen für eine signifikante Modellverletzung. Einzelne Aufgaben, wie z.B. Aufgabe 19 („Wer ist das?" – Foto von Daimler-Vorstand Dieter

Zetsche) weichen sowohl bei der Aufteilung nach dem Rohwert als auch bei der Aufteilung nach Geschlecht von der Winkelhalbierenden ab. Diese Aufgabe war für Frauen und Personen mit insgesamt niedrigerer Allgemeinbildung schwieriger als für Männer und Personen mit insgesamt höherer Allgemeinbildung. Andere Aufgaben weichen bei der Aufteilung nach dem Rohwert stärker von der Winkelhalbierenden ab als bei der Aufteilung nach Geschlecht.

Wichtig ist dabei zu beachten, dass ein Unterschied in der geschätzten Schwierigkeit der Aufgaben nicht durch die unterschiedlichen Fähigkeiten der Personen verursacht wird: Auch wenn Personen mit höherer Fähigkeit (und entsprechend höherem Rohwert) eine höhere Wahrscheinlichkeit haben, eine Aufgabe richtig zu beantworten, sollte die Schwierigkeit der Aufgabe im Verhältnis zu den anderen Aufgaben für alle Personen gleich sein. Eine Abweichung einer Aufgabe von der Winkelhalbierenden in Abbildung 4 (links) entspricht hingegen der Situation zweier sich überschneidender Aufgaben wie in Abbildung 1 (rechts): Aufgabe 23 („Was bedeutet das sechseckige Bio-Siegel?" – es dürfen keine chemisch-synthetischen Pflanzenschutzmittel verwenden werden) ist z.B. im Vergleich zu den anderen Aufgaben für Personen mit höherer Fähigkeit schwieriger zu beantworten als für Personen mit niedrigerer Fähigkeit – und für Frauen leichter als für Männer.

Abbildung 4: Vergleich der geschätzten Aufgaben-Schwierigkeiten bei der Aufteilung nach Fähigkeit (links) und Geschlecht (rechts).



Die Nummern in der Abbildung stehen für die Nummern der Aufgaben im Fragebogen.

Um zu überprüfen, ob die Annahmen des Rasch-Modells auch für andere Gruppen verletzt sind, wird im folgenden Abschnitt ein neues statistisches Verfahren vorgestellt, bei dem mehrere Merkmale gleichzeitig daraufhin untersucht werden können, ob die resultierenden Gruppen von Probanden signifikante Unterschiede in den Aufgaben-Schwierigkeiten und damit DIF aufweisen. Nur wenn sich die Aufgaben-Parameter bei keiner möglichen Aufteilung der Stich-

probe unterscheiden, können Personen mit unterschiedlichem soziodemografischen Umfeld und Bildungshintergrund objektiv und fair miteinander verglichen werden.

3.4 Das neue statistische Verfahren

Für den grafischen Modelltest im letzten Abschnitt haben wir verschiedene Personen-Gruppen vorgegeben und geprüft, ob sich die Schwierigkeiten von einzelnen Aufgaben für diese Gruppen unterscheiden. Im Folgenden möchten wir ein neues statistisches Verfahren vorstellen, das automatisch Gruppen von Personen findet, deren Aufgaben-Parameter sich unterscheiden. Diese Gruppen können sich insbesondere auch aus mehreren Merkmalen wie Geschlecht, Alter und Bildungsstand zusammensetzen. Der Vorteil dieses explorativen Vorgehens ist, dass – sofern alle relevanten Merkmale erfasst wurden – keine Personengruppen übersehen werden, in denen sich die Aufgaben-Parameter unterscheiden, und die deshalb nicht fair mit demselben Test beurteilt werden können.

Das Verfahren basiert auf Tests aus der Wirtschaftsstatistik, die z.B. dazu verwendet werden können, systematische Veränderungen, sog. Strukturbrüche, in Aktienkursen o.ä. nachzuweisen. Dieselben statistischen Tests lassen sich nun auch dazu verwenden, zu überprüfen, ob sich z.B. die Schwierigkeit einer bestimmten Aufgabe systematisch für jüngere und ältere Probanden unterscheidet. Liegt zwischen den Ausprägungen eines Merkmals wie dem Alter, Bildungsstand oder Geschlecht ein Strukturbruch in den Schwierigkeits-Parametern einzelner Aufgaben vor, werden die Personen anhand des Merkmals in zwei Gruppen aufgeteilt, für die dann getrennte Rasch-Modelle mit den entsprechenden unterschiedlichen Aufgaben-Parametern angepasst werden, wie z.B. in Abbildung 5. Daran erkennt man sowohl, dass ein gemeinsames Modell nicht geeignet ist, die unterschiedlichen Personengruppen fair zu vergleichen, als auch, welche Aufgaben sich in welchen Gruppen besonders unterscheiden. Das Modell stellt eine Weiterentwicklung des Ansatzes von Strobl, Wickelmaier, and Zeileis (2010) für das Rasch-Modell dar und steht in der frei zugänglichen Statistik-Software R (R Development Core Team 2010) im Erweiterungspaket psychotree (Zeileis, Strobl, Wickelmaier und Kopf 2009) zur Verfügung.

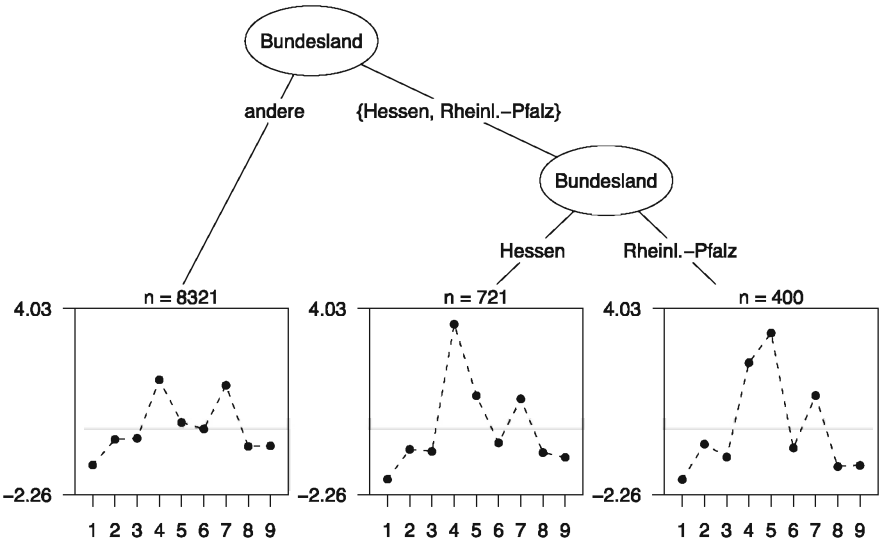
Das Prinzip des neuen statistischen Verfahrens und seine Anwendung zur Diagnose von DIF soll anhand eines einfachen Beispiels erklärt werden: In dem von uns betrachteten Fragebogen kommen zwei Fragen vor, bei denen Probanden, die in einem bestimmten Bundesland leben oder aufgewachsen sind, im Vorteil sind („Wo liegt Hessen?" – Auswahl auf der Deutschlandkarte, „Wie heißt die Landeshauptstadt von Rheinland-Pfalz?" – Mainz). Der Wohnort der Probanden wurde zwar nicht direkt abgefragt – aber das Bundesland, in dem ggf. das Abitur gemacht wurde. Dabei stellt sich heraus, dass Probanden, die in Rheinland-Pfalz Abitur gemacht haben, signifikant besser abschneiden als Probanden, die in einem anderen deutschen Bundesland Abitur gemacht haben.

Dieses Ergebnis kann zwei mögliche Ursachen haben: Es wäre möglich, dass Abiturienten aus Rheinland-Pfalz tatsächlich eine bessere Allgemeinbildung haben als alle anderen Probanden. Wahrscheinlicher ist aber, dass die Abiturienten aus Rheinland-Pfalz durch die Konstruktion des Fragebogens, der eine Frage zur Landeshauptstadt von Rheinland-Pfalz selbst und eine Frage zur Lage des Nachbar-Bundeslandes Hessen enthält, im Vorteil sind. Deshalb überprüfen wir im Folgenden, ob die zentrale Annahme des Rasch-Modells, dass eine Aufga-

be im Vergleich zu den übrigen Aufgaben nicht für manche Personen einfacher zu lösen sein darf als für andere, erfüllt ist.

Wir betrachten zunächst nur die ersten neun Aufgaben des Fragebogens zum Themengebiet Politik. Die vierte Aufgabe ist: „Wo liegt Hessen?", die fünfte Aufgabe ist: „Wie heißt die Landeshauptstadt von Rheinland-Pfalz?". Wir überprüfen nun, ob es einen Strukturbruch in den Schwierigkeits-Parametern einzelner Aufgaben gibt, wenn man als Einflussgröße das Bundesland, in dem das Abitur gemacht wurde, wählt. Das Ergebnis ist in Abbildung 5 dargestellt: Abiturienten aus Hessen fällt eindeutig die vierte Aufgabe am leichtesten, Abiturienten aus Rheinland-Pfalz fällt sowohl die vierte als auch die fünfte Aufgabe deutlich leichter als die anderen Aufgaben.

Abbildung 5: Unterschiedliche Aufgaben-Parameter nach Bundesland.



Aufgabe 4: „Wo liegt Hessen?", Aufgabe 5: „Wie heißt die Landeshauptstadt von Rheinland-Pfalz?". Aufgaben die weit oben liegen sind für die jeweilige Gruppe besonders einfach zu lösen. Die Nummern an den x-Achsen der Abbildung stehen für die Nummern der Aufgaben im Fragebogen. Die Zahl n gibt die Anzahl der Personen in der jeweiligen Gruppe an.

Die Unterschiede in den Aufgaben-Parametern sprechen dafür, dass Abiturienten aus Rheinland-Pfalz nicht wirklich eine höhere Allgemeinbildung haben, sondern dass das Ergebnis durch DIF verzerrt wird. Dieser Verdacht lässt sich leicht überprüfen: Streicht man die beiden Bundesland-spezifischen Aufgaben aus dem Test, und schätzt die Allgemeinbildung der Probanden erneut, schneiden Probanden, die in Rheinland-Pfalz Abitur gemacht haben, tatsächlich nicht mehr besser ab als die übrigen Abiturienten. Bevor man also einen Test sinnvoll für Gruppenvergleiche einsetzen kann, muss er zunächst auf DIF überprüft werden.

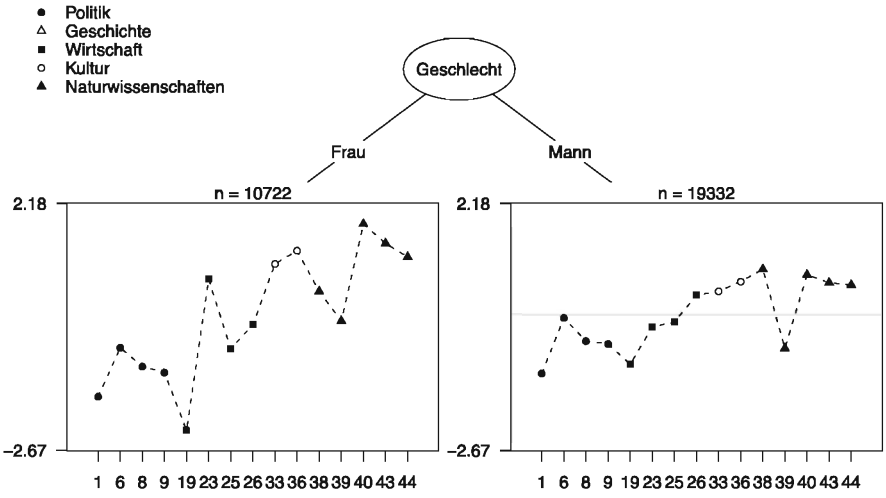
4. Untersuchung der Geschlechterunterschiede

Das neue Verfahren soll nun verwendet werden, um DIF als mögliche Ursache für die beobachteten Geschlechterunterschiede im Studentenpisa zu untersuchen. Zunächst wird deshalb nur das Merkmal Geschlecht verwendet, um zu überprüfen, ob ein Strukturbruch in den Aufgaben-Parametern vorliegt. Dabei stellt sich heraus, dass sich die Aufgaben-Parameter von Frauen und Männern für viele der 45 Aufgaben systematisch unterscheiden.

4.1 Unterschiede in den Aufgaben-Parametern

Die gravierendsten Unterschiede zwischen Frauen und Männern finden sich in den Bereichen Politik und Naturwissenschaften. Betrachtet man nur die 15 Aufgaben mit den stärksten Unterschieden, wie in Abbildung 6, sind Aufgaben aus dem Bereich Geschichte hingegen gar nicht vertreten, da die Schwierigkeiten der meisten Aufgaben aus diesem Themenbereich zwar insgesamt relativ hoch sind, aber sich kaum zwischen Frauen und Männern unterscheiden. (Die in Abbildung 6 dargestellten Aufgaben mit den stärksten Unterschieden zwischen Frauen und Männern sind natürlich auch diejenigen, die in Abbildung 4 (rechts) am stärksten von der Winkelhalbierenden abweichen.)

Abbildung 6: Unterschiedliche Aufgaben-Parameter nach Geschlecht.



Aufgaben, die weit oben liegen, sind für die jeweilige Gruppe besonders einfach zu lösen. Die Nummern an den x-Achsen der Abbildung stehen für die Nummern der Aufgaben im Fragebogen. Die Zahl n gibt die Anzahl der Personen in der jeweiligen Gruppe an.

Abbildung 6 belegt, dass es sowohl Themenbereiche gibt, die Frauen leichter fallen (Kultur, aber auch Naturwissenschaften), als auch Themenbereiche, die Frauen vergleichsweise schwerer fallen als Männern (Politik und Wirtschaft). Aufgaben, die für Frauen leichter – und für Männer vergleichsweise schwierig – zu lösen sind, sind z.B. die Aufgabe 23 („Was bedeutet das

grüne, sechseckige Bio-Siegel?" – es dürfen keine chemisch-synthetischen Pflanzenschutzmittel verwenden werden) und Aufgabe 40 („Was wird auch als ‚Trisomie 21' bezeichnet?" – das Down-Syndrom). Im Vergleich zu den anderen Aufgaben besonders schwierig war für Frauen die Aufgabe 19 („Wer ist das?" – Foto von Daimler-Vorstand Dieter Zetsche).

Für die im SPIEGEL-Artikel erwähnte Aufgabe 33 („Wie heißt der Bestsellerroman von Daniel Kehlmann?" – Die Vermessung der Welt), die im Hinblick auf die Rohdaten von Frauen häufiger richtig beantwortet wurde als von Männern, wird in Abbildung 6 deutlich, dass sich insbesondere das Verhältnis der Schwierigkeit dieser Aufgabe im Vergleich zu den übrigen Aufgaben für Frauen und Männer unterscheidet: Frauen fällt diese Aufgabe – und auch Aufgabe 36 („Welche der folgenden Opern ist NICHT von Mozart" – Aida) – leichter als Aufgabe 38 („Wofür wird Ultraschall NICHT genutzt?" – Radio); Männern hingegen fallen die beiden Aufgaben aus dem Bereich Kultur schwerer. (Die absolute Schwierigkeit der Aufgaben kann man in dieser Darstellung nicht direkt vergleichen, da die Aufgaben-Parameter für Frauen und Männer getrennt geschätzt und normiert wurden.)

4.2 *Wie viele Aufgaben sind betroffen?*

Werden die geschätzten Aufgabenparameter auf dieselbe Skala gebracht, z.B. indem der Parameter der Aufgabe mit dem geringsten DIF als gemeinsamer Nullpunkt festgelegt wird, kann man die Differenzen zwischen den geschätzten Aufgabenparametern für Frauen und Männer genauer untersuchen: Insgesamt weisen 30 der 45 Aufgaben ein statistisch signifikantes DIF zwischen Frauen und Männern auf. In Tabelle 1 ist die Häufigkeit von Aufgaben ohne signifikantes DIF sowie von Aufgaben mit signifikantem DIF, bei dem Frauen bzw. Männer die Aufgabe als leichter empfinden, getrennt nach den fünf Themengebieten dargestellt. Daran lässt sich wiederum ablesen, dass Männer insbesondere bei Aufgaben aus den Bereichen Politik und Wirtschaft im Vorteil sind, während Frauen sowohl bei Fragen zu Kultur, als auch bei Fragen zu Naturwissenschaften im Vorteil sind. Insgesamt sind von den 30 Aufgaben mit signifikantem DIF jeweils 15 für Frauen und 15 für Männer leichter zu beantworten.

Tabelle 1: Aufgaben mit und ohne DIF nach Themengebieten

	kein DIF	leichter für Frauen	leichter für Männer
Politik	3	1	5
Geschichte	4	2	3
Wirtschaft	2	2	5
Kultur	4	5	0
Naturwissenschaft	2	5	2

Auffallend an den Ergebnissen ist, dass – abgesehen vom Themengebiet Geschichte, das Frauen und Männern gleichermaßen schwer fällt – die Unterschiede in den Aufgaben-Schwierigkeiten der Themengebiete Wirtschaft und Politik vs. Kultur und Naturwissenschaften dafür sprechen, dass die Allgemeinbildung nicht auf einer eindimensionalen Skala abgebildet wer-

den kann, sondern dass unterschiedliche Personengruppen auf den einzelnen Dimensionen des Fragebogens unterschiedlich abschneiden. Bei dieser mehrdimensionalen Betrachtung gibt es also neben den Bereichen, in denen die Männer im Vergleich vorne liegen, auch Gebiete, in denen die Frauen klar überlegen sind.

4.3 Ist DIF für die gefundenen Geschlechterunterschiede verantwortlich?

Im Gegensatz zu dem einfachen Beispiel der Bundesland-spezifischen Aufgaben im Themenbereich Politik, bei dem der scheinbare Unterschied in der Allgemeinbildung von Abiturienten aus Rheinland-Pfalz und den übrigen Bundesländern verschwindet, wenn die Aufgaben mit DIF aus dem Test entfernt werden, wird der Unterschied zwischen Männern und Frauen durch die Entfernung der DIF-Aufgaben nur etwas abgeschwächt (ohne Abbildung). Der Grund hierfür ist, dass die Aufgaben, die für Frauen besonders schwierig zu beantworten sind, auch absolut betrachtet eine höhere Schwierigkeit haben. Entsprechend wirkt sich die Entfernung der 30 Aufgaben mit DIF – obwohl mit jeweils 15 Aufgaben die Anzahl der Aufgaben, die für Frauen bzw. Männer schwieriger zu beantworten sind, gleich ist – leicht zugunsten der Frauen aus. Diese Ergebnisse decken sich auch mit denen, die Bertling et al. in einer Untersuchung mit anderen statistischen Verfahren erzielt haben und in einem anderen Beitrag in diesem Band darstellen.

Allerdings bleibt der Unterschied in der Allgemeinbildung zugunsten der Männer auch nach Entfernung der Aufgaben mit DIF statistisch signifikant. Der im Studentenpisa gefundene Geschlechterunterschied kann also nicht allein durch DIF erklärt werden.

Mögliche Ursachen für den bestehenden Geschlechterunterschied werden in den vorangegangenen Kapiteln diskutiert. Wichtig ist aber natürlich bei der Diskussion um Geschlechterunterschiede zu berücksichtigen, dass das Geschlecht (insbesondere das biologische Geschlecht) nie als alleinige oder gar kausale Ursache für beobachtete Unterschiede interpretiert werden kann, da zwischen den Geschlechtern auch Unterschiede in der Sozialisation und anderen möglichen Einflussgrößen bestehen, die nicht oder nur schwer messbar – aber dennoch untrennbar mit dem erhobenen Merkmal Geschlecht verwoben – sind.

4.4 Weitere Ergebnisse

Abschließend betrachten wir alle Personen aus der Studentenpisa-Stichprobe gemeinsam, um zu überprüfen, ob noch weitere Personengruppen unterschiedliche Aufgaben-Parameter aufweisen. Aus allen zur Verfügung stehenden Merkmalen zum soziodemografischen und Bildungshintergrund der Probanden zeigen sich signifikante Unterschiede zwischen mehreren Gruppen, die in Abbildung 7 und Abbildung 8 als Baum-Struktur dargestellt sind. An dieser Darstellung erkennt man gut, dass sich das neue Verfahren (im Gegensatz zu den üblichen Methoden zur Überprüfung von DIF, wie dem graphischen Modelltest, bei dem je zwei Gruppen zum Vergleich fest vorgegeben werden müssen) auch einsetzen lässt, um Gruppen von Personen mit unterschiedlichen Aufgaben-Schwierigkeiten zu identifizieren, die sich durch eine komplexe Kombination von mehreren Merkmalen zusammensetzen.

Abbildung 7: Linker Teil des Baumes: Unterschiedliche Aufgaben-Parameter in der Gesamtstichprobe.

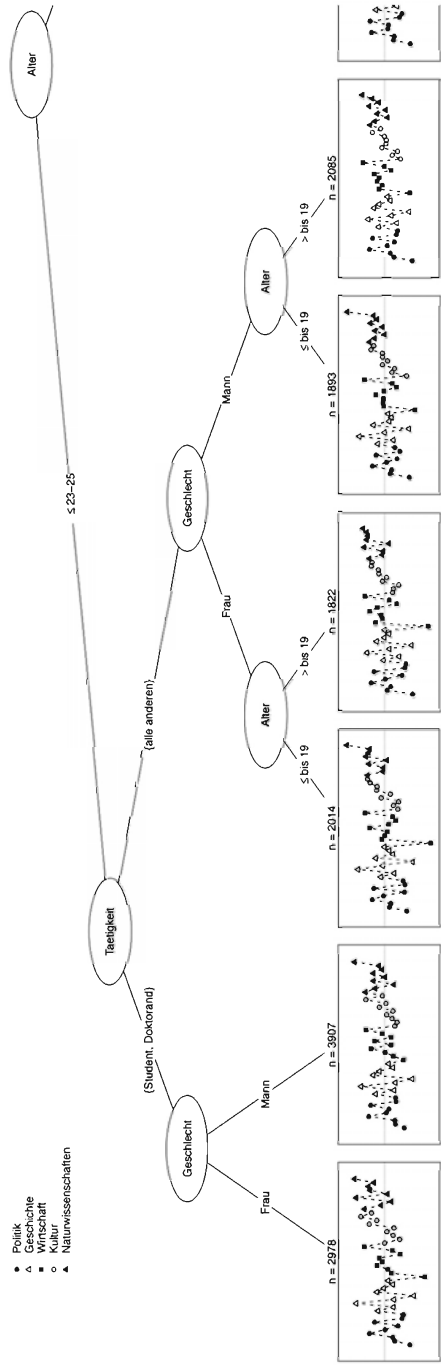
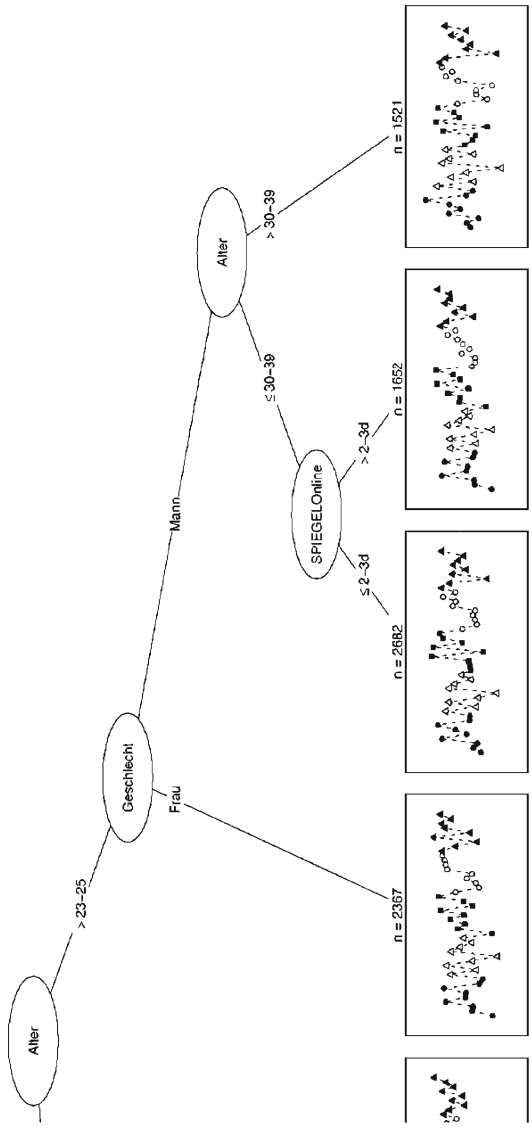


Abbildung 8: Rechter Teil des Baumes: Unterschiedliche Aufgaben-Parameter in der Gesamtstichprobe.



Aufgaben, die weit oben liegen, sind für die jeweilige Gruppe besonders einfach zu lösen. Die Zahl n gibt die Anzahl der Personen in der jeweiligen Gruppe an. Von allen zur Auswahl stehenden soziodemographischen Variablen hat nur das Alter, das Geschlecht, die Tätigkeit und das Lesen von SPIEGEL ONLINE einen Einfluss auf die Aufgaben-Schwierigkeit.

Betrachtet man z.B. die Gruppe der Frauen im Alter bis 25 Jahren, die Studentin oder Doktorandin sind (ganz links in Abbildung 7), stellen sich die Aufgaben 12 („Welche Herrschaftsform verbindet man mit dem französischen König Ludwig XIV?" – Absolutismus), 33 („Wie heißt der Bestsellerroman von Daniel Kehlmann?" – Die Vermessung der Welt), 35 („In welcher Stadt steht dieses Gebäude?" – Bild des Eiffelturms in Paris) und 40 („Was wird auch als ‚Trisomie 21' bezeichnet?" – das Down-Syndrom) als im Vergleich zu den übrigen Aufgaben verhältnismäßig leichter, die Aufgabe 25 („Welcher deutsche Konzern hat den britischen Autobauer Rolls-Royce übernommen?" – BMW) hingegen als verhältnismäßig schwieriger heraus.

Betrachtet man hingegen die Gruppe der Männer im Alter über 39 Jahren (ganz rechts in Abbildung 8), stellen sich die Aufgaben 7 („Welches EU-Organ wird 2009 von den Bürgern der EU-Mitgliedstaaten gewählt?" – das Europäische Parlament), 10 („Die römische Seeherrschaft wurde begründet..." – durch die Niederwerfung Karthagos) und 18 („Wodurch konnte Mao-Zedong seine Macht in China ausbauen?" – Langer Marsch) als im Vergleich zu den übrigen Aufgaben verhältnismäßig leichter, die Aufgaben 29 („Was haben diese vier Gebäude gemeinsam?" – alle vier wurden vom selben Architekten entworfen), 32 („In welcher Fernsehserie wurde der US-Präsident lange von einem afroamerikanischen Darsteller gespielt?" – in der Serie ‚24') und 39 („Welche Sinneszellen im menschlichen Auge sind für das Farbsehen verantwortlich?" – Zapfen) hingegen als verhältnismäßig schwieriger heraus.

Weitere Aufgaben, die sich für einzelne Gruppen als besonders leicht oder schwierig zu lösen erwiesen haben, sind: die Aufgabe 22 („Was ist ein CEO?" – Chief Executive Officer), die für Männer zwischen 23 und 39, die mindestens 2-3 mal täglich SPIEGEL ONLINE nutzen (der dritte von rechts in Abbildung 8), vergleichsweise einfach ist, die Aufgabe 28 („Von welchem Maler stammt dieses Bild?" – Bild von Andy Warhol), die für Männer bis 19 Jahren, die nicht Student oder Doktorand sind (der fünfte von links in Abbildung 7), vergleichsweise schwierig ist, sowie die Aufgabe 45 („Was ist die Summe der Innenwinkel eines Dreiecks?" – 180 Grad), die über alle Personen hinweg die leichteste Aufgabe darstellt (vgl. Abbildung 2), aber für Frauen und Männer bis zum Alter von 19 Jahren, die nicht Studenten oder Doktoranden sind (der dritte und fünfte von links in Abbildung 7), im Vergleich zu den übrigen Aufgaben noch einfacher zu lösen ist – was natürlich daran liegt, dass die allermeisten Befragten unter 20 noch Schüler sind, bei denen die Trigonometrie zum Schulstoff gehört.

5. Zusammenfassung der Ergebnisse und Fazit

Unterschiede in den Aufgaben-Parametern von verschiedenen Personengruppen können, wie am Beispiel der Bundesland-spezifischen Aufgaben dargestellt, zu systematischen Verzerrungen von Testergebnissen führen. Deshalb wurde in diesem Kapitel überprüft, ob der im Studentenpisa gefundene Geschlechterunterschied durch DIF verursacht worden sein könnte.

Die Ergebnisse zeigen, dass 30 der 45 Aufgaben des hier betrachteten Fragebogens signifikante Unterschiede in den Schwierigkeiten für Frauen und Männer aufweisen: Während Männer viele Aufgaben aus den Bereichen Wirtschaft und Politik einfacher finden, tun sich Frauen mit vielen Aufgaben aus dem Bereich Kultur, aber auch mit Aufgaben aus den Naturwissenschaften leichter.

Obwohl die Aufgaben, die von Frauen als besonders schwierig empfunden wurden, auch absolut gesehen besonders schwierig sind, führt die Entfernung der Aufgaben mit DIF nur zu einer leichten Abschwächung des Geschlechterunterschiedes. Der signifikante Unterschied in

der Allgemeinbildung zugunsten der Männer ist also kein rein methodisches Artefakt. Mögliche Ursachen dieses Geschlechterunterschiedes werden in vorangegangenen Kapiteln diskutiert; das Geschlecht kann aber nicht als kausale Ursache von Leistungsunterschieden interpretiert, sondern nur im Zusammenhang mit Unterschieden in der Sozialisation und anderen Einflussgrößen betrachtet werden.

Die unterschiedlichen Aufgaben-Schwierigkeiten für Frauen und Männer in den unterschiedlichen Themenbereichen des Studentenpisa spricht zudem dafür, dass die Allgemeinbildung sich nicht als eindimensionales Konstrukt messen lässt. Auf den unterschiedlichen Dimensionen einer mehrdimensionalen Skala würden sich hingegen sowohl Bereiche auszeichnen, in denen Männer besser abschneiden, als auch Bereiche, in denen Frauen besser gebildet sind. Zu diesen Bereichen gehört unseren Ergebnissen nach nicht nur der Bereich Kultur – an dem sich im SPIEGEL-Artikel die Debatte entzündet hatte, ob z.B. das Wissen über Literatur oder Musik „genauso wichtig" ist wie das Wissen über Politik – sondern auch der Bereich Naturwissenschaften.

Neben den Unterschieden zwischen Frauen und Männern zeigen sich in unserer weiteren Auswertung auch Unterschiede zwischen Personen unterschiedlichen Alters, unterschiedlicher Tätigkeiten und unterschiedlicher Mediennutzung.

Zusammenfassend zeigen die Ergebnisse unserer statistischen Analyse also, dass es nicht möglich ist, aus den Ergebnissen des Studentenpisa einfache Schlussfolgerungen über die Überlegenheit einzelner Personengruppen abzuleiten. Ein komplexer Test wie dieser erfordert eine differenziertere Betrachtung aller möglichen Quellen von Gruppenunterschieden, inklusive derer, die in der Konstruktion des Tests selbst begründet sind.

Insbesondere bedeutet das Vorliegen von DIF in einem Rasch-Modell auch, dass die Betrachtung der Rohwerte der Personen (also der Anzahl der richtig gelösten Aufgaben, wie im SPIEGEL-Artikel) – genauso wie die Betrachtung der aus einem gemeinsamen Modell geschätzten Allgemeinbildung – zu falschen Aussagen führen kann, weil die Aufgaben nicht für alle Personen anhand einer einzelnen Dimension „Allgemeinbildung" angeordnet werden können.

Literatur

Asendorpf J, Wenderlein M (2009). Darf man mit IQ-Tests Ethnien oder Geschlechter vergleichen? *Gehirn & Geist*, 5, 14-17.

Ben-Shakhar G, Sinai Y (1991). Gender Differences in Multiple-Choice Tests: The Role of Differential Guessing Tendencies. *Journal of Educational Measurement*, 28(1), 23-35.

Grütz D (2004). Der geschlechtsspezifische Zugriff auf Lesestrategien – Ergebnisse einer Untersuchung im Rahmen unterrichtsdidaktischer Forschung. *Linguistik online*, 21(4).

Hedges L, Nowell A (1995). Sex Differences in Mental Health Test Scores, Variability, and Numbers of High-Scoring Individuals. *Science*, 269(5220), 41-45.

Jahnke-Klein S (2005). Chancengleichheit für Mädchen und Jungen im mathematisch-naturwissenschaftlichen Unterricht. In F Hellmich (Hrsg.), *„Lehren und Lernen nach IGLU – Grundschulunterricht heute,"* Oldenburg: Didaktisches Zentrum (diz).

Mair P, Hatzinger R (2007). "Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R." Journal of Statistical Software, 20(9). URL http://www.jstatsoft.org/v20/i09.

Rasch G (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* The University of Chicago Press, Chicago. Neuauflage 1980.

R Development Core Team (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schnurr EM (2007). Frauen sind auch nur Männer. ZEIT Wissen, 1. URL http://www.zeit.de/zeit-wissen/2007/01/Titel-Frauen-Maenner.

Simpson R, Oliver J (1985). Attitude toward Science and Achievement Motivation Profiles of Male and Female Science Students in Grades Six through Ten. *Science Education*, 69(4), 511-526.

Strobl C, Wickelmaier F, Zeileis A (2010). Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning. *Journal of Educational and Behavioral Statistics*. (Im Druck.).

Trenkamp O (2009). Mädchen fürchten Mathe, Jungs schwächeln beim Lesen. SPIEGEL ONLINE Schul-SPIEGEL, 26.05. URL http://www.spiegel.de/schulspiegel/wissen/0,1518,626879,00.html.

Verbeet M (2009). Dramatische Differenz. SPIEGEL, 21, 34-40.

Woike B, Bender M, Besner N (2008). "Implicit motivational states influence memory: Evidence for motive by state-dependent learning in personality." Journal of Research in Personality, 43(1), 39-48.

Zeileis A, Strobl C, Wickelmaier F, Kopf J (2010). psychotree: Recursive Partitioning Based on Psychometric Models. R package version 0.10-0, URL http://CRAN.R-project.org/package=psychotree.

Danksagung

Carolin Strobl wird im Rahmen des Projekts STR1142/1-1 („Methoden zur Berücksichtigung von Subjekt-Kovariablen in IRT-Modellen") von der Deutschen Forschungsgemeinschaft gefördert. Außerdem möchten wir Reinhold Hatzinger für die wichtigen Anregungen durch unsere Gespräche und sein R-Paket eRm (Mair und Hatzinger 2009) danken.