Andreas VOHNS, Klagenfurt

Zermelo, Rasch, Schrödinger: Ein stoffdidaktischer Zugang zur probabilistischen Modellierung mathematischer Leistung

1. Einführung und Motivation

Probabilistische Testmodelle, allen voran das Rasch-Modell, spielen in der Erfassung mathematischer Leistungen in den letzten gut 15 Jahren eine zunehmende Rolle, initiiert durch die großen internationalen Schulleistungsvergleiche TIMSS und PISA. Dabei wird dem Modellcharakter der eingesetzten Test- und Messmodelle meinem Eindruck nach nur relativ wenig Aufmerksamkeit gewidmet. Betrachtet man das Zugänglicher-Machen von Mathematik als ein Grundanliegen stoffdidaktischen Arbeitens und weist dem Herausarbeiten des „mathematischen Kerns der Sache“ (Kirsch, 1977) dafür eine wesentliche Rolle zu, so kann man hinsichtlich des Rasch-Modells stoffdidaktischen Forschungsbedarf sehen: In der Diskussion um Vor- und Nachteile der Verwendung des Rasch-Modells für mathematische Leistungserfassung wird bei Befürwortern wie Gegnern nicht immer gründlich zwischen dem Modell selbst und davon unabhängigen Erweiterungen unterschieden. Der Einfluss des Modells auf die Leistungsmessung wird dann u. U. sowohl über- als auch unterschätzt.

Im Vortrag (und abgekürzt in diesem Beitrag) wird versucht, das Rasch-Modell in einer anderen Einkleidung (Modellierung der Spielstärke im Schach) mathematisch besser zugänglich zu machen und diesen Kontext dann mit dem eigentlich interessierenden Kontext der Modellierung von (mathematischen) Leistungsdaten zu kontrastieren. Besondere Bedeutung nimmt dabei die Frage ein, welche „Erfüllungsnormen“ (Schreiber, 1980) das Rasch-Modell an (mathematische) Tests und deren Bearbeitung anlegt. Eines kritischen Blicks wird die „spezifische Objektivität“ der Rasch-Skala gewürdigt, die bisweilen auch unter der leicht missverständlichen Bezeichnung „Stichprobenunabhängigkeit“ firmiert. Abschließend wird diskutiert, inwiefern die Tendenz zur Bildung von „Subskalen“ nicht auf eine gewisse kognitive Dissonanz in der Verwendung des (strikt eindimensionalen) Rasch-Modells für inhaltlich eher breite Konzepte wie „mathematische Kompetenz am Ende einer Schulstufe“ hinweist.

2. Einkleidung: Zermelos abgebrochenes Schach-Turnier

Bereits mehr als dreißig Jahre vor Georg Raschs Nutzung eines probabilistischen Modells für die Konstruktion und Auswertung von Tests nutzte Ernst Zermelo (1929) ein sehr ähnliches Verfahren, um im Falle eines vorzeitig abgebrochenen Schach-Turniers eine faire Bewertung der beteiligten

In F. Caluori, H. Linneweber-Lammerskitten & C. Streit (Hrsg.), *Beiträge zum Mathematikunterricht 2015*. Münster: WTM-Verlag

Spieler zu ermitteln (die resultierende Bewertung der Spielstärke ist eng mit den Elo-Zahlen verwandt). Um die Analogie zur Leistungsmessung zu verdeutlichen, präsentiere ich es in leicht abgewandelter Form:

Ein Schach-Verein mit 14 Mitgliedern will seine 5 spielstärksten Mitglieder zu einem internationalen Turnier schicken. Die Mitglieder unterteilen sich in 5 „Meister“ (waren im letzten Jahr beim internationalen Turnier) und 9 „Stümper“ (waren im letzten Jahr nicht dabei). Ein Vorauswahlturnier soll entscheiden. Es findet in zwei Runden statt. Erste Runde: Jeder Meister tritt gegen jeden Stümper an. Zweite Runde: Alle Meister treten gegeneinander an, ebenso alle Stümper gegeneinander. Zum internationalen Turnier fahren diejenigen Spieler, die insgesamt die größte Anzahl an Siegen für sich verbuchen können.

Das Turnier muss nach der ersten Runde abgebrochen werden. Wie kann man nun entscheiden, wer insgesamt am besten war? Zermelos Vorschlag läuft darauf hinaus, unter den Meistern weiterhin nach Anzahl der Siege (gegen 0, 1, …, 9 Stümper) zu sortieren, ebenso unter den Stümpern (gegen 0, 1, …, 4 Meister). Ein Meister gilt ferner dann besser als eine Gruppe von Stümpern (mit gleicher Anzahl von Siegen), wenn er mehr als 50% der Partien gegen diese Gruppe gewonnen hat. Ebenso gilt ein Stümper als besser als eine Gruppe von Meistern (mit gleicher Anzahl von Siegen), wenn er mehr als 50% der Partien gewonnen hat. Diese Regeln lösen das Sortierproblem aber nur zum Teil, weil es zu Intransitivitäten kommen kann (Beispiel: Meister A ist besser als Stümper B, der besser als Meister C ist, der wiederum besser ist als Stümper D. Stümper D ist aber besser als Meister A).

Die probabilistische Modellierung ersetzt nun empirisch aufgetretene relative Häufigkeiten von Siegen (in Gruppen gleich starker Spieler) durch ML-geschätzte (gemäß einer speziellen logistischen Funktion) geglättete Gewinnwahrscheinlichkeiten, die stets zu transitiven Ordnungen auf der Vereinigungsmenge von Meistern und Stümpern führt und die Ordnungen gemäß tatsächlichen Siegen auf den beiden Teilmengen selbst respektiert.

3. Anwendung: Leistungsmessung

Hier treten nicht Meister gegen Stümper an, sondern Personen gegen Items. Ein Item „gewinnt“ gegen eine Person, wenn es falsch (bzw. nicht zustimmend) beantwortet wird, eine Person gewinnt gegen ein Item, wenn es korrekt (bzw. zustimmend) beantwortet wird. Das „Turnier“ ist hier notwendig unvollständig: Personen können nicht direkt gegen Personen antreten, Items keine Items bearbeiten. Man kann dennoch die Regeln von oben adaptieren, um Items und Personen wieder auf eine gemeinsame Skala an-

zuordnen: Ein Item ist „schwieriger“ als ein anderes Item, wenn es insgesamt weniger oft korrekt gelöst wurde, eine Person „fähiger“ als eine andere, wenn sie mehr Items korrekt gelöst hat. Eine Gruppe von gleichfähigen Personen wird höher als ein Item eingeschätzt, wenn der Anteil korrekter Lösungen größer als 50% ist, umgekehrt ein Item besser als eine Personengruppe, wenn der Lösungsanteil kleiner als 50% ist. Auch hier können Intransitivitäten auftreten (Beispiel: Item A ist zu schwierig für Person B, die Item C beherrscht, das wiederum zu schwierig ist für Person D. Person D beherrscht aber Item A).

Die probabilistische Modellierung ersetzt aufgetretene relative Lösungshäufigkeiten (in Gruppen gleichfähiger Personen) durch ML-geschätzte (gemäß einer bestimmten logistischen Funktion, dem sog. Rasch-Modell) geglättete Lösungswahrscheinlichkeiten, die stets zu einer transitiven Ordnung auf der Vereinigungsmenge von Items und Personen führt und die Ordnung gemäß Anzahlen korrekter Lösungen auf den beiden Teilmengen selbst respektiert (zum Schätzverfahren vgl. Rost, 1996).

4. Erfüllungsnormen und Konsequenzen

Zentrale, wenn nicht einzige Erfüllungsnorm der Rasch-Modellierung ist die Eindimensionalität der Messung. Für die Fähigkeitsschätzung im Rasch-Modell ist (im Falle des Ein-Matrix-Designs, also: alle Personen bearbeiten sämtliche Items) ausschließlich die Anzahl der Items entscheidend, die korrekt gelöst wurden, nicht aber welche Items. Für die Schwierigkeitsschätzung eines Items ist einzig entscheidend, wie viele Personen es nicht korrekt bearbeitet haben, nicht welche. Das Modell wird die realen Daten daher umso besser approximieren, desto homogener das Lösungsverhalten ist, d.h.: Das Modell setzt voraus bzw. passt dann gut, wenn a) bei einem Test Personen, die die gleiche Anzahl von Items korrekt bearbeitet haben, stets auch in etwa dieselben Items korrekt bearbeitet haben und b) eine Gruppe A von Personen, die mehr Items als eine andere Gruppe B korrekt bearbeitet, möglichst viele der Items auch korrekt bearbeitet, die die schwächere Gruppe B korrekt bearbeitet. Relative Stärken von Gruppen (z.B. eine Hälfte der 5 Items lösenden Personen löst eher die Items 1-5 korrekt, die zweite eher die Items 6-10) können mit dem Rasch-Modell nicht erklärt werden, sie sind Residuen (zufällige Abweichungen und/oder Anzeichen dafür, dass ein mehrdimensionales Konstrukt gemessen wird).

Während in der Psychometrie ein wichtiger Einsatzzweck der Rasch-Modellierung in der Überprüfung der Eindimensionalität von Konstrukten besteht, wird diese Annahme bei mathematischen Leistungsmessungen regelmäßig aus testpragmatischen Gründen schlicht unterstellt und auch zur

Itemselektion in der Pilotierung von Tests herangezogen. Entscheidendes Motiv für die pragmatische Unterstellung der eindimensionalen Modellierbarkeit mathematischer Leistung ist vor allem die Möglichkeit, eine Rasch-Modellierung in Multi-Matrix-Designs einzusetzen, d.h. dort, wo gerade nicht allen Personen sämtliche Items administriert werden können oder sollen. Bearbeiten verschiedene Gruppen von Personen verschiedene Testhefte, so kann bei ausreichend großen Schnittmengen der Items zwischen den Testheften und unter Voraussetzung der Passung des Rasch-Modells ein Vergleich aller Personen und aller Aufgaben auf einer gemeinsamen Skala immer noch vorgenommen werden. Dabei können die Testhefte sogar gezielt unterschiedlich schwer sein (um etwa verschiedene Jahrgänge miteinander zu vergleichen).

5. Stichproben(un)abhängigkeit und Lösungswahrscheinlichkeit

Der Vortrag schließt mit einer Betrachtung der sog. „Stichprobenunabhängigkeit“ (= spezifische Objektivität). Diese Eigenschaft kommt dem Modell (approximativ dann dem Datensatz) zu und besagt im Kern, dass bei Bildung von Teiltests (nur ein Teil der getesteten Aufgaben wird berücksichtigt) die Rangreihenfolge der Personen unverändert bleiben muss. Es wird dann argumentiert, dass ein Konzept wie „Lösungswahrscheinlichkeit“ in diesem Modell immer von der (Gesamt-)Stichprobe abhängig ist, insofern es seine Zufälligkeit zunächst nur aus der Ziehung einer Person aus einer Gruppe gleich viele Items korrekt lösender Personen bezieht. Inwiefern eine individuelle Interpretation des Wahrscheinlichkeitswerts sinnvoll ist, kann in Frage gestellt werden (hier wird ein Bezug zu „Schrödingers Katze“ hergestellt, vgl. Wainer, 2010).

Bezüglich der spezifischen Objektivität wird auch diskutiert, inwiefern die Bildung von Subskalen bzgl. eines Tests, der bereits einer Itemselektion bzgl. einer Gesamtskala unterzogen wurde, eine vertretbare statistische Praxis darstellt.

Literatur

Kirsch, A. (1977). Aspekte des Vereinfachens im Mathematikunterricht. *Didaktik der Mathematik, 5*, 87–101.

Rost, J. (1996). *Lehrbuch Testtheorie Testkonstruktion*. Bern: Hans Huber.

Schreiber, A. (1980). Idealisierungsprozesse — ihr logisches Verständnis und ihre didaktische Funktion. *Journal für Mathematik-Didaktik, 1 (1-2)*, 42-61.

Wainer, H. (2010). Schrödinger’s Cat and the Conception of Probability in Item Response Theory. *Chance*, *23 (19)*, 53-56.

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift, 29 (1)*, 436–460.

Foliensatz unter: http://de.slideshare.net/andreasvohns/zermelo-schrdinger-rasch