Heiner Willenberg

Forschungsbasierte Einführung in das Raschmodell für die deutschdidaktische Empirie

Auf der Basis des Programms Winsteps

Thema: Mangel an Kompetenzforschung

Man fühlt sich etwas an Friedrich Karl Waechters Karikatur erinnert: "Wahrscheinlich guckt wieder kein Schwein" (1978), wenn es um den Kompetenzbegriff geht. Seit über zwei Jahrzehnten stellen die großen Forschungsprojekte angefangen mit PISA und weiter mit IGLU, DESI und den Arbeiten des IQB vielfältige Kompetenzstufungen vor (s. die Dissertation von Thomas Canz 2020), die das Raschmodell benutzen und die damit zum Standard geworden sind - nur in den deutschdidaktischen Forschungen, die unabhängig von diesen Großprojekten arbeiten, taucht diese Verfahrensweise praktisch nicht auf (s. a. Klieme et al. 2010). In der Bibliographie pedocs findet sich unter den Stichwörtern "Raschmodell" oder "raschskaliert" insgesamt 78 Eintragungen in allen Fächern und in der allgemeinen Pädagogik (Anfang 2023). Der Deutschdidaktik lässt sich allenfalls eine Handvoll von Publikationen zurechnen, die zudem öfters ihre Statistiken im Hintergrund verstecken.[1] Ein mustergültiges Projekt aus der Schreibdidaktik wird zur Anregung am Schluss des Berichts vorgestellt: Blatt, Ramge und Voss (2008).

Zwar hatte Dorothee Wieser (2019) als Kompromiss angeboten, wir müssten nur wissen, worum es bei diversen statistischen Verfahren geht - die genaue Analyse könnten wir uns von Empirikern rechnen lassen. Es scheint aber nicht, dass sie damit eine offene Tür gefunden hat - z.B. erscheint in dem umfassenden und intensiven dreibändigen Werk von Jan M. Boelmann (2018f.) zu den Grundlagen empirischer Forschung dieses Verfahren mit keinem Wort: Wo es kein Handbuchwissen gibt, da entstehen auch keine Fragen.

Einführung in das Raschmodell mit dem Computerprogramm Winsteps

So soll hier eine Einführung in das Raschmodell vorgestellt werden, die sich auf Beispiele und Leistungen englischsprachiger universitärer Untersuchungen stützt, d. h. es werden keine Großprojekte herangezogen, sondern lediglich Untersuchungen im Rahmen von 75 bis zu einigen hundert Probanden. In der Nomenklatur gilt die Item Response Theorie (IRT) als übergeordnete Testrichtung, innerhalb derer das Raschmodell (Rasch 1960/1980) als das klare Verfahren für die Eindimensionalität der zu erfassenden latenten Fähigkeiten benutzt wird (ausführlich Kelava/ Moosbrugger 2020).

Verbunden mit dieser Einführung ist ein Einblick in das Computerprogramm *Winsteps* (Linacre 1998 ff), das in der englischsprachigen pädagogischen Forschung dominiert und das den großen Vorteil hat, mit einer Eingabemaske zu arbeiten und keine Syntaxzeilen zu verlangen. Es ist in allen vorgestellten Beispielen das benutzte grundlegende Hilfsmittel. Um eine eventuelle Benutzung zu erleichtern, werden hier die entsprechenden Hinweise in Kursivschrift angegeben.
Winsteps für Windows kam 1998 auf den Markt und wird seitdem von seinem Autor John Michael Linacre bis in die Gegenwart ständig verbessert, aber in seinem grundsätzlichen Aufbau nicht verändert. Die Prozedur ist einfach, mit einem gewissen Haken.

Datentypen /Control File

Das Programm liest verschiedene Datentypen: Excel, SPSS, Text-Tabs und SAS. Danach muss dem System mitgeteilt werden, welche Informationen zu den Testpersonen gehören und welche Itemwerte sind - eigentlich selbstverständlich -, wenn die Bezeichnungen nicht per Markierung und Verschieben (drag and drop) in die entsprechenden Zeilen des erschienenen ersten Formulars geschoben werden müssten. Verständliche kurze Einführungen dazu bei YouTube: First Step on Using Winsteps (o.J.) oder im Buch von Rita Green (2013) Kapitel 10.
Wenn dann die Grundlagen erstellt sind *(Control File)*, kann die Rechnung losgehen. Es erscheint als erstes eine "*Summary*", in der u.a. die Reliabilitäten stehen und auch die Anzahl der Iterationen mit ihrer erreichten Güte. Und danach kann man über 40 Tafeln aufrufen, die alle möglichen Auskünfte über die Items und die Testpersonen geben, u.a. auch die Wright-Map. Gesondert findet man eine Vielzahl von Grafiken. Es ist klar, schreiben Boone et al. (2014: 460 f.), dass man normalerweise nur eine kleine Auswahl dieser Ergebnislisten braucht, im vorliegenden Bericht sind lediglich sechs verwendet bzw. erwähnt worden.
Für erste Versuche mit dem Programm kann man sich die kostenlose Version Ministeps herunterladen, die auf 25 Items und 50 Personen begrenzt ist. Wie eingangs erwähnt ist Winsteps das am meisten benutzte Instrument in der englischsprachigen pädagogischen Forschung, die inzwischen auch weite Teile Asiens umfasst.

Das Raschmodell

Der grundlegende Ansatz des Raschmodells unterstützt Leistungsmessungen und dabei besonders die Erarbeitung von Kompetenzstufen, dafür werden die vier grundsätzlichen Annahmen vorgestellt.

a Iterationen: Das zentrale Verfahren der Iterationen besteht darin, Personen- und Item-Werte in mehrfachen Durchgängen gegeneinander abzugleichen. Dafür werden die Werte in eine Gesamtmatrix gesetzt und das System vergleicht sie vielfach, um die beste Passung zwischen den Resultaten der beiden "Partner" zu erreichen: Sind es wirklich die Personen mit den hohen Rohwerten, die auch die eher komplizierten Fragen bewältigt haben und halten sich die Abweichungen (die Residuen) in einem akzeptablen Rahmen? Winsteps benutzt dafür die geläufigste Methode, die Joint-Maximum Likelihood-Equation, mit der Abkürzung JMLE. Das Wort "joint" bezeichnet die Verbindung der Personen mit den Items. Und "Likelihood" heißt soviel wie Plausibilität.

b Die Eindimensionalität: Das Modell verlangt grundsätzlich, dass nur eine Fähigkeit erfasst wird, z. B. die Lesekompetenz. Die Eindimensionalität muss zunächst durch die Validität des Konstrukts angestrebt werden, also dadurch dass die Aufgaben z.B. auf allen Ebenen die Lesefähigkeit anpeilen und nicht z. B. inhaltliches Wissen. Rechnerisch wird sie durch die gute Passung von Fähigkeiten und Anforderungen gezeigt, die ja durch die Iterationen demonstriert wurde. Fehlt diese Passung bei Probanden oder Items, dann entstehen schlechte Fitwerte, stimmt sie, sehen wir gute Fitwerte, die weiter unten (Kapitel 3) als wichtige Prüfkategorien präsentiert werden. Insgesamt sind hier die Residuen im Spiel, die mitteilen, wie stark die Ergebnisse von einer erwarteten Ideallinie abweichen, der gebräuchliche Wert RMSE wäre im Idealfall 0, kleinere Werte etwa unter 1 zeigen eine gute oder ausreichende Genauigkeit an.

c Wahrscheinlichkeitsrechnungen sind für Geisteswissenschaftler eher ungewöhnlich. Sie werden hier eingesetzt, und zwar so, dass bei den Personenwerten nicht die Rohwerte in die Berechnung eingehen, sondern es wird die Wahrscheinlichkeiten kalkuliert, mit denen eine Person die Aufgaben verschiedener Schwierigkeiten bewältigt. Zu den Berechnungen der Wahrscheinlichkeiten: Wenn jemand z.B. 80% gelöste Aufgaben hat, kann man rechnen: 80 richtige zu 20 nicht gelöste: 80/20 - ergibt 4. Und wenn eine Person nur 20 % der Aufgaben gelöst hätte: 20/80 wäre 0,25 (s. Bond u.a. 2015, S. 26f.).

Da die Schwierigkeiten der Items in dieselbe Richtung wie die Fähigkeiten der Personen laufen sollen, muss dafür die Fragerichtung, die bei den Personen angewandt wird, für die Items umgedreht werden, nämlich wie wahrscheinlich ist jetzt ihre *Nicht*lösung und somit wird eine Gegenwahrscheinlichkeit berechnet (Kelava/Moosbrugger: 374 f.). Eine Aufgabe, die von 80% der Personen nicht gelöst wird, sondern nur von 20%, bekommt in dieser Gegenwahrscheinlichkeit die Notierung 80/20 d.h. 4 und den Logit (siehe Punkt d) von 1.38, die Aufgabe ist also schwer.

d Logits: Die Statistiker setzen die Werte der Wahrscheinlichkeiten in ihre Logarithmen um: 1 wird zu 0; 4 wird zu 1.38; 0,25 wird zu - 1,38. Diese sogenannten Logits gruppieren sich um den Mittelwert von 0 - die fähigeren Personen liegen im positiven Bereich darüber, der in etwa bis + 4

geht, die schwächeren unterhalb, circa bis -4. Die Aufgaben folgen derselben Ausrichtung mit der Gegenwahrscheinlichkeit. Der "Trick" mit dem Logits führt dazu, dass es keine Gerade der Resultate gibt, sondern eine leicht geschwungene Kurve, die Item-Charakteristik-Kurve (englisch ICC), die die Ergebnisse an den Rändern sinnvollerweise auseinanderzieht (s. Bond 2020, 24f.)

Die sechs Eigenarten und Vorteile des Raschmodells

1 Kompetenzstufen

Es ist allgemein bekannt, dass die großen Projekte ihre Modelle in vier- bis fünfstufigen Kompetenzstufen präsentieren, die von einfachen grundlegenden Anforderungen bis zu dem Anspruch an komplexes Verstehen reichen.[2] Essentiell dafür ist es, die Aufgaben vorher in eine theoriebegründete Hierarchie zu bringen. Die Ergebnisse zeigen dann, ob diese Hierarchie auch von den Probanden und deren festgestellten Fähigkeiten bestätigt wurde.

Für einen ersten Einblick wird hier ein von Rita Green (2013) veröffentlichtes Leseprojekt "Reading 2" herangezogen: 195 Schüler/innen der 11. Klasse sollten 20 Leseaufgaben lösen, die Kodierung war auf 0/1 festgelegt, also dichotom (122 ff).

Zunächst hier die etwas rudimentäre Beschreibung der vier Theoriestufen, die auch durch den Kennzeichnungen teilweise unterschieden werden sollen.

< lokale Lektüre "scanning" (Item 1-5)

< explizite Hauptideen erfassen (Item 6-10) *(mit einem < versehen)*

< implizite Hauptideen (11-15) (*mit einem * versehen)*

< schwierige Wörter (16-20) (Green: 124)

Hier treffen wir auf den Roten Faden dieses Berichts - die detaillierte Aufgaben-beschreibung ist auch für Rita Green nicht wichtig, obwohl ihr Buch in allen anderen Bereichen außerordentlich klar ist - auch auf Nachfrage waren die genauen Aufgabenformulierungen nicht zu bekommen. Dafür sind die berichteten Resultate aber interessant, sowohl im positiven wie im kritischen Bereich.

Winsteps: Item Entry (Reihenfolge der Items) / Item measure (Items nach Ergebnissen)

In der vereinfachten Tabelle sehen wir, dass die Messwerte (JMLE Measures) für die Stufe der schwierigen Wörter (links) alle im oberen Teil der Tabelle liegen und die diejenigen Lektüre alle im leichten unteren Teil, vier von ihnen unter dem Mittelwert von 0 und sie überlappen sich nicht. Die beiden anderen Gruppen, die sich mit den Hauptideen befassen, streuen dagegen breit.

Die fünf Items mit schwierigen Wörtern

Item-nummer	Rohwert	JMLE Measures
17	70	1.63
20	78	1.34
19	91	.88
18	96	.71
16	110	.21

Die Werte für die globalen Lektüren

Itemnummer	Rohwert	JMLE Measures
<8	8	8
*11**	*11*	*11*
<10	10	10
*12**	*96*	*.71*
<9	100	.57

Die fünf Items für die lokale Lektüre

Item-nummer	Roh-werte	JMLE Measure
1	117	.04
5	144	-1.12
2	153	-1.53
3	153	-1.53
4	176	-2.96

Fortsetzung der globalen Lektüre

Itemnummer	Rohwert	JMLE Measures
*13**	*113*	*.10*
*15**	*121*	*-.19*
<7	123	-.27
<6	142	-1.03
*14**	*148*	*-1.29*

Tab. 1: Ergebnisse eines Lesetests ("Reading 2") mit zwei passenden und zwei divergenten Kompetenzstufen (Green 2013: 171)

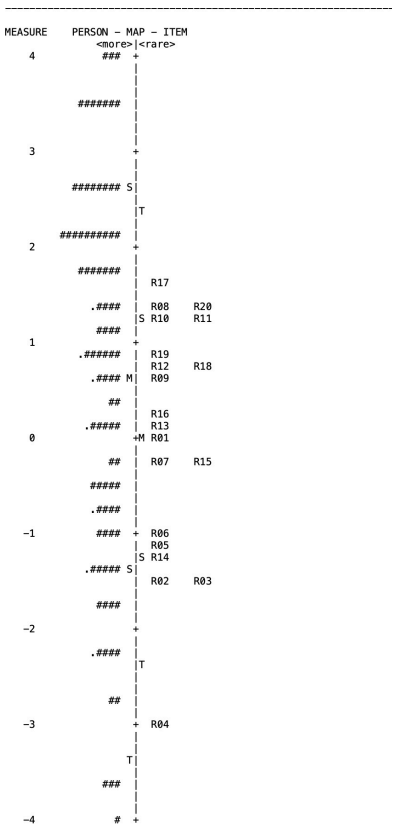
Die zehn Items für die globale Lektüre erstrecken sich fast über die gesamte Größe der Messtabelle, Item 8 ist mit 1.34 das zweitschwerste und Item 14 mit -1.29 das viertleichteste. Und auch die Unterteilung in zwei globale Gruppen, wie im Original vorgeschlagen, bringt nichts, denn die Markierungen, die hier kursiv oder fett gesetzt sind, zeigen ein ziemliches Durcheinander der beiden Gruppen an. So wird man als Fazit formulieren können: Zwei der vermuteten Kompetenzen sind klar unterschieden, die beiden im Raschmodell nicht - sie sollten noch einmal theoretisch überarbeitet und umformuliert werden.

2 Graphische Übersicht über die Ergebnisse: Wrighttafeln oder der Röntgenblick

Item: Wrightmap

Der amerikanische Statistiker Benjamin D. Wright, der von Anfang an das Raschmodell propagierte, hat eine graphische Übersicht eingeführt, in der man die Passung der Aufgaben und der Probanden gut erkennen kann - als Exempel die Wrighttafel des Projekts "Reading 2" von Rita

Green. Auf der folgenden Seite: Links zeigen die Kreuze die Verteilung der Personen an, von oben nach unten durch ihre Fähigkeiten geordnet, auf der rechten Seite die Items, auch entsprechend ihrer Schwierigkeiten. Ganz links die Skala von - 4 (unten) über den Mittelpunkt von 0 bis + 4.
Wrighttafeln vermitteln auch größeren Projekten eine gute Übersicht darüber, wo sich die einzelnen Teile nach ihren Resultaten befinden und wo es Lücken in den Anforderungen gibt. In einem südkoreanische Sprachtest für Englisch lagen vier Subtests vor: Hörverstehen, Leseverstehen, Grammatik und Wortschatz. Die Tafel dokumentiert: Die Probanden sind gleichmäßig verteilt, wenn auch deutlich oberhalb des Nullpunktes. In zwei der vier Testteile hingegen gibt es unterschiedlich große Lücken, die der Autor mit den eingefügten Kennzeichnungen sichtbar hervorhebt (Leaper, 2010): Hörverstehen hat zu wenig leichte Aufgaben, Grammatik deutlich zu wenig schwere. Quasi ein Röntgenbild des Projekts.



Man sieht deutlich, dass links im Bereich von +1,5 bis + 4 vier Zeilen

mit den Kreuzen für die Probanden stehen, die ca. 70 Personen reprä-sentieren. Rechts gibt es in dieser Höhe keine Items. Diese Kombination zeigt, der Test war für einen größeren Teil der Schüler/innen zu leicht. Für eine Forschungsgruppe, die sich mit dem Lesetest beschäftigt, bringt diese Tafel die Einsicht, dass die Aufgaben verändert werden werden sollten.

Abb. 1: Wrighttafel des Lesemodells Reading 4 (Green 2013: 163)

3 Fitwerte - die passenden Items

Ein zentraler Kontrollwert in allen Rasch- und IRT-Untersuchungen sind die Fitwerte der Items, mit denen festgestellt wird, ob sich das Ergebnis einer Aufgabe in die postulierte Einlinigkeit einpasst oder ob es deutlich darüberhinaus geht. Rechnerisch werden die Residualwerte benutzt, also die Abweichungen von der nie erreichbaren Idealpassung zwischen Personen und Aufgaben.

Zwei Kategorien werden mitgeteilt: Infit und Outfit, jeweils mit ihren quadrierten Mittelwerten und mit standardisierten Daten. Ein zu großer Infit heißt, das Item und seine Ergebnisse diskriminieren zu wenig, ein zu großer Outfit zeigt, das Item war zu schwer, es befindet sich außerhalb der erwarteten Verteilung. Der quadrierte Mittelwert sollte im Rahmen von 0,5 bis 1,5 liegen, sein standardisierter Wert, der hier herangezogen wird, müsste sich zwischen -2,0 und +2,0 befinden. Linacre ermutigt: "There is no one "correct" statistic. Use the one you find most useful" (2012: 200). Zur Illustration sei eine Untersuchung vorgestellt, die keine Leistungsdaten erhebet, sondern die Einschätzungen von Personen erfragen, und dabei polytome Varianten benutzen, bei denen mehrstufige Zustimmungen oder Ablehnungen verwendet werden. Bei allen Befragungen zu den Einschätzungen ist es wichtig zu wissen: Fragen mit häufiger und leichter Zustimmung rangieren im Bereich der Leichtigkeit, also in den Minuswerten unterhalb des Mittelwerts, seltene Akzeptanz ähnelt den seltenen Leistungsergebnissen, wird also oben in den positiven Werten lokalisiert.
Das Projekt: Es wurden 75 deutsche Biologie-Lehrkräfte in der Ausbildung danach gefragt, wie sie ihre Wirksamkeit im gewählten Unterrichtsfach einschätzen. Das benutzte Instrument ist der STEBI-Fragebogen: Science Teaching Efficacy Belief Instrument (Enochs/Riggs 1990), ausführlich dargestellt in Boone et al. 2014. Die Fragen ergeben positive aber auch zögerliche Aussagen:
Häufige Zustimmungen: Q2: "Ich werde kontinuierlich bessere Wege finden, Biologie[3] zu unterrichten." (-2.49) und Q 22: "Wenn ich Biologie unterrichte, begrüße ich die Fragen der Schüler/innen üblicherweise" (-1.83). Beispiel für eine zögerliche Antwort Q 5: "Ich kenne die notwendigen Schritte, um biologische Konzepte effektiv zu unterrichten." (1.16)

Winsteps: Item Entry, dort: Infit /Outfit

Eine Item fällt auf (Q3), es hat den standardisierten Outfit von +3,1, also über der Grenze von 2,0, und das obwohl es in der Ergebnisskala der Zustimmungen mit 0.22 nahe am Mittelwert liegt und dadurch völlig unauffällig ist. Wenn man sich die Formulierung der Frage ansieht, findet man: "Auch wenn ich mir sehr viel Mühe gebe, werde ich Biologie nicht so gut unterrichten wie die meisten (anderen) Fächer." Im Gegensatz zu allen anderen Fragen wird von den jungen Lehrkräften hier ein Bezugnahme zu ihrem zweiten oder dritten Fach verlangt - und damit wird eine Kategorie außerhalb des befragten Hauptthemas angesprochen. Deshalb liegen wohl die Resultate nicht nahe an der gradlinigen Einschätzung zur Biologie, sondern streuen weit umher zwischen leichten und seltenen Zustimmungen. Das Projekt offenbart eine Problematik der Antworten, die aus den reinen Rohwerten bzw. aus den Logits/Measures nicht hervorgehen würden, und die sich nur im Raschmodell entdecken lassen. Angefügt werden soll noch die Beobachtung, dass in Berichten über

Leistungsmessungen derartige Ausreißer selten berichtet werden, weil sie wahrscheinlich schon nach der Pilotierungsphase ausgeschlossen wurden.

4 Die genauen und die ungenauen Distraktoren

Oft erscheint es sinnvoll, Multiple-choice-Aufgaben (Mehrfachwahl-Aufgaben) in einem Projekt einzusetzen. Der Vorteil liegt darin, dass man es vermeidet, die Probanden zur eigenen Formulierung aufzufordern. Die Auswahl der exakt zutreffenden Antwort im Verbund mit ähnlichen setzt bei den Testpersonen durchaus ein genaues Wissen oder eine genaue Interpretation voraus. Man umgeht damit das komplexe Thema, offene Antworten interpretieren zu müssen. Anspruchsvoll ist es aber, die Formulierungen der falschen Varianten relativ nahe an die richtige Antwort zu bringen, aber eben auch nicht zu nahe. In diesem Feld ermöglicht das Raschmodell eine gute Einschätzung der Distraktoren. Es ist zur Auswertung nötig, Winsteps mitzuteilen, welcher der Distraktoren der richtige ist.
Rita Green hat in ihrem Lesetest "Reading 4" vier Antwortmöglichkeiten untergebracht, die mit Buchstabenkürzeln a, b, c, d versehen wurden, deren Lösungsschlüssel so aussieht: abbcacbcdaaa.

Winsteps: Distraktoren. Control file - key Eintrag

Item-num-mer	Code	Wert	Roh-wert	Durchschnitt-liche Fähigkeit
1	b	0	3	-1.59
	d	0	3	-.74
	c	0	1	-.64
	a	1	**95**	**.45**
3	x	0	2	-1.19
	d	0	21	-.28
	c	0	22	-.23
	a	0	16	-.19
	b	1	**41**	**1.25**
6	d	0	15	.13
	b	0	32	.35
	x	0	6	.44
	a	0	18	.64
	c	1	**30**	**.25***

Tab. 2: Die genauen Werte für die Distraktoren (Green, 2013) (Item Entry 14.3)

Bei Winsteps wird im control file folgender Eintrag gemacht: *; GROUPS = 0 ; Partial Credit model: in case items have different rating scales. CODES = abcdx ; matches the data. KEY = abbcacbcdaa (*genau so mit Lücken!) und oberhalb der Personenlabel per Hand eingefügt, nachdem der control file zunächst aufgerufen worden ist (Green 2013: 186).

Es ergibt sich in dieser Tafel eine hilfreiche Ergebnisliste, die jedem Distraktor neben den Prozenten seiner Zustimmung noch die durchschnittliche Fähigkeit derjenigen Personen zuweist, die den Distraktor gewählt haben. Erwartet wird, dass falsche Ankreuzungen mit niedrigeren Personenwerten einhergehen als bei den richtigen. Im folgenden Ausdruck ist die richtige Variante in der untersten Zeile des Items zu sehen und jeweils mit einem Sternchen* versehen. Damit haben - ein sehr gutes Mittel in der Hand, diese Distraktoren zu überarbeiten. Wie üblich zeigt die Statistikerin Green aber nicht an, um welche sprachlichen Nuancen es sich handelt - immerhin ist die Analysemethode für sprachlich-literarische Projekte relativ leicht zu benutzen.

5 Differente Funktion eines Items (DIF) - gibt es doch Untergruppen?

Winsteps: Item DIF, between/within

Die DIF-Analyse bei Winsteps ist einfach: Man muss nur nach der ersten "Summary" die Tafel DIF anklicken und bekommt dann die Auswahl derjenigen Parameter angeboten, die in den Probandendaten eingetragen worden waren.

Die strikte Interpretation der Einlinigkeit im Raschmodell würde etwaige Abweichungen bei Personengruppen nicht tolerieren. John Linacre, der Winsteps konstruiert hat, formuliert flexibler und mit ihm viele andere: Die Differenten Funktionen verletzten das Gebot der Homogenität nicht, wenn sich die Itemkurven der beiden Gruppen ähneln. Und Michelle Raquel (2019: 107) führt vor, wie die Itemkurven im Idealfall nur verschoben aber nicht verändert sind: Die linke und die rechte Kurve zeigen die beiden unterschiedlichen Gruppen an, in der Mitte die Gesamtgruppe.

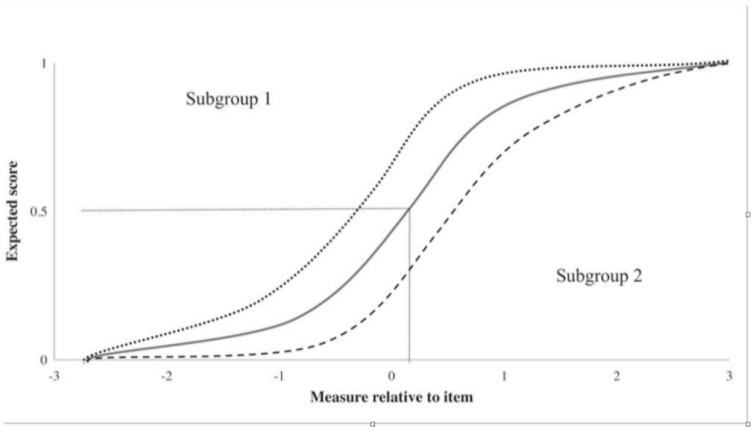


Abb. 2: Die idealisierten Item-Charakeristikkurven bei einem Item mit gleichlaufender DIF (Raquel 2021: 107). In der Waagerechten die Fähigkeitsmaße für die Personen des Tests, in der Senkrechten die Skala der Lösungwahrscheinlichkeiten. Eine Person der Subgroup 1 (obere Linie) mit seinem Maß von -0.5 löst die Aufgabe mit 50%-iger Wahrscheinlichkeit, jemand der Subgroup 2 mit -0.5 (rechte Linie) mit ca. 15%-iger Wahrscheinlichkeit.

John Linacre hat eine aufschlussreiche Untersuchung an 75 Kindern und deren Interesse an der Natur publiziert (2012), bei der es 25 Fragen gab. Einige seien erwähnt: Wie interessant ist es für dich, einen Zoo zu besuchen, auf ein Picknick zu gehen, Vögel zu beobachten oder auch Flaschen und Dosen draußen aufzusammeln?

KID CLASS	OBSERVATIONS COUNT	AVERAGE	BASELINE EXPECT	MEASURE	DIF SCORE	DIF MEASURE	DIF SIZE	DIF S.E.	DIF t	ACT Prob.	Number	Name
F	18	1.61	1.59	-.40	.02	-.48	-.09	.49	-.17	.8638	1	Watch birds
M	56	1.39	1.40	-.40	-.01	-.40	.00	.23	.00	1.000	1	Watch birds
F	18	.44	.73	2.42	-.29	3.47	1.05	.48	2.17	.0458	5	Find bottles
M	56	.48	.39	2.42	.09	2.09	-.33	.24	-1.34	.1871	5	Find bottles

Tab. 3: Linacre: DIF bezüglich dem Interesse an der Natur. Der blaue Kasten zeigt in der dritten Zeile den Logit-Unterschied von 1.05. Es war den Mädchen deutlich unangenehmer, Flaschen zu sammeln als den Jungen (find bottles).

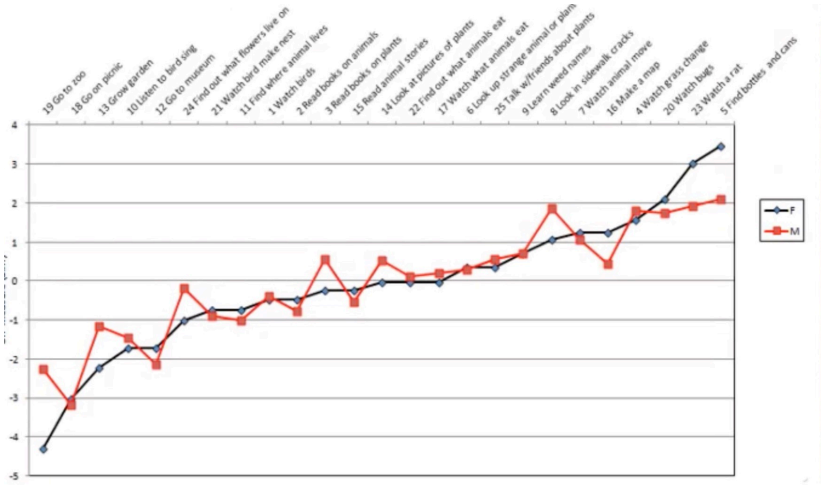


Abb. 3: Die Zustimmungswerte von Jungen und Mädchen zu 25 Fragen, die sich auf ihr Interessen an der Natur beziehen, die dunkle Linie steht für die Mädchen, die rote für die Jungen. (Linacre 2018).

Grundsätzlich - und das ist wichtig - stimmten die Linien der 25 Items von Jungen und Mädchen überein. Bei einigen aber gab es sichtbare Unterschiede: Die Jungen bequemten sich eher der unbeliebten Aufgabe an, Flaschen und Dosen aufzusammeln (find bottles) mit einem Unterschied im Logit von 1.05, was als gravierend gilt. Die beiden anderen Fragen zum Picknick und zur Vogelbeobachtung wurden von beiden Geschlechtern völlig gleich bewertet.

Die schwarze Linie steht für die Präferenzen der Mädchen, die helle für die der Jungen, wie üblich finden sich die niedrigen Logits links für große Zustimmung, die rechten höheren für eine zögerliche Akzeptanz. Ganz links der Zoobesuch, ganz rechts das Flaschensammeln. Diese Art, Einlinigkeit und milde Abweichungen zu demonstrierten, gehört für Linacre und andere in den Rahmen des Raschmodells!

Noch ein Hinweis auf ein anderes Projekt. R. J. de Ayala (2022) beschreibt einen amerikanischen Vokabeltest, bei dem neun Wörter aus dem Slang der Schwarzen benutzt wurden und die Frage war, ob die weißen Amerikaner etwas davon verstehen. Die Antwort, die sich aus den Ergebnissen für die zwei Untergruppen speist, lautet: teils ja, teils kaum. Diese Thematik scheint für Deutschland nicht anwendbar zu sein, aber der Wortschatztest von DESI zeigte doch, dass die Jungen fast in allen Sparten schwächere Leistungen erbrachten außer dort, wo bei einem Bahnhofsbild eine Handvoll von technischen Wörtern genannt werden sollte (Willenberg 2007). So könnte man bei künftigen Untersuchungen vermutete Präferenzen der Jungen einbauen und die Items auf die Differente Funktion prüfen.

6 Partial credit - die teilweise Anerkennung der Leistung

Bei schwierigen Leseaufgaben ist es interessant, offene Aufgaben zu entwerfen und die Antworten der Getesteten in einer üblichen Viererskala von "nicht erfasst" bis "vollständig erfasst" durch die Testergruppe einzuordnen. Dass hier das Problem der Rater-übereinstimmung entsteht, lassen wir aus. Es sei darauf hingewiesen, dass John Linacre dafür ein umfassendes Computerprogramm namens "Facets" veröffentlicht hat, das leider mit einer zehnzeiligen Syntax arbeitet. Dazu gibt es bei Rita Green mit dem Kapitel 13 die beste Einführung.
An dieser Stelle wird eine andere ermutigende Fähigkeit von Winsteps gezeigt, mit der es möglich ist, dichotome Daten mit den polytomen der partial-credit-Modells in einem homogenen Ansatz zu verbinden.
Es war sehr schwierig, ein Projekt aus der Sprachforschung mit partial-credit-Items zu finden. Fan und Bond beschreiben einen Test zum Hörverstehen der englischen Sprache, publiziert von der chinesische Universität Fudan unter dem Titel "Fudan English Test". Dort finden sich vier Items in der Abteilung "spot dictation", bei denen das Diktat einen Stop einlegt und die Probanden aufgefordert werden, in die schriftliche Testlücke ihr Verständnis des Wortes einzutragen. Den Probanden werden ihre "Guthaben" von 1-4 nach der Qualität ihrer Antworten gegeben, schreiben Fan und Bond. Es handelt sich in der folgenden Liste um die Items 5-8 (siehe Entry number). Man erkennt sofort, dass die Rohwerte (total score) dafür deutlich höher sind als bei den Nachbarn,

trotzdem kann das System diese Items in die Einlinigkeit einordnen, weil ihm vorher mitgeteilt wurde, dass es sich um Partial-Credit-Aufgaben handelt, deren Werte von 0-4 reichen.

Winsteps: Im Control File muss dann eingetragen werden:
"GROUPS = 0; GROUPS = 0 ; Partial Credit model: in case items have different rating scales CODES = 01234 ; matches the data." (Der Code 0 steht für keine oder eine völlig falsche Antwort)

TABLE 14.1 Winsteps specification file Unidimens ZOU468WS.TXTo Dec 13 2022 17:45lsx
INPUT: 106 PERSON 25 ITEM REPORTED: 106 PERSON 25 ITEM 58 CATS WINSTEPS 5.3.2.0

PERSON: REAL SEP.: 1.80 REL.: .76 ... ITEM: REAL SEP.: 3.54 REL.: .93

ITEM STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	JMLE MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM	G
1	94	106	-1.36	.32	.84	-.65	.59	-1.15	.44	.28	89.5	89.2	SD1	0
2	91	106	-1.07	.30	.95	-.20	.84	-.42	.35	.30	86.7	86.4	SD2	0
3	76	106	-.08	.23	1.11	1.06	1.25	1.35	.22	.35	75.2	74.0	SD3	0
4	70	106	.23	.22	.82	-2.16	.72	-2.06	.53	.36	75.2	70.1	SD4	0
5	293	106	-1.41	.19	.76	-1.01	.80	-.22	.49	.42	83.8	81.3	SD5	0
6	131	106	.18	.18	1.13	1.04	1.13	.98	.35	.45	63.8	64.6	SD6	0
7	231	106	.91	.10	.98	-.12	.96	-.22	.64	.63	37.1	39.4	SD7	0
8	169	106	.89	.13	.76	-2.09	.74	-2.14	.69	.56	51.4	47.6	SD8	0
9	76	106	-.08	.23	1.15	1.43	1.18	1.00	.21	.35	71.4	74.0	C1	0
10	86	106	-.68	.26	1.05	.39	1.23	.91	.24	.32	82.9	81.9	C2	0

Tab. 4: Vier Items (5-8), die eine Kodierung von 0-4 haben (Fan/Bond: 121 bzw. mit den verfügbaren Daten der Routledge Website nachgerechnet)

Man erkennt eine übliche Verteilung der Schwierigkeiten von - 1.41 bis 0.91. Die punktbiseriale Einordnung der vier Items liegt klar über den empfehlenswerten .30. Der Outfit von Item 8 ist mit -2.14 nicht befriedigend. Wichtig ist aber, dass sich hier interpretative Aspekte in das quantitative Grundmodell integrieren lassen.

Ein umfassendes Projekt mit dem Raschmodell

Die Forschungsgruppe um Inge Blatt, Gesa Ramm und Andreas Voss (2008) hat eine differenzierte, mehrstufige Skala zur Einschätzung von schriftlichen Schülertexten entwickelt und nach der Pilotierung im Haupttest an 431 Jugendlichen aus sechsten Klassen getestet.
Sie stellten die Aufgabe, sich für eine Brieffreundschaft zu "bewerben" und zwar auf eine Anzeige hin, in der ein deutsches Mädchen, das zur Zeit in Hawai lebt, einen Briefpartner oder eine Briefpartnerin sucht. Sie werde wohl bald nach Deutschland zurückkehren.
Die Gruppe bezog sich in der Gestaltung der Aufgaben auf die Schreibtheorie von Carl Bereiter (1980) und teilte die erwarteten Fähigkeiten in grundlegende, dabei auch formale, und in erweiterte

ein. Zu den erweiterten gehört der Bezug auf das Thema, auf die Adressatin und auch eine gewisse Darstellung der eigenen Person. Am Ende gibt es 15 Beurteilungs-Items, die alle dichotom ausgewertet wurden.

Zur empirischen Absicherung setzte die Gruppe das Raschmodell ein, das mit *ConQuest* berechnet wurde, ein mächtiges Programm, wie es heißt, allerdings eines mit ausgiebigen Syntaxketten.

Welche Aspekte aus dem Raschmodell wurden benutzt?

1 Die Fitwerte: Fünf Items wurden wegen Infitproblemen aussortiert (weitere nach der Pilotierung und Schwierigkeiten mit dem Codebuch). So blieben die erwähnten 15 Items übrig.

2 Wrighttafel: Sie zeigt sehr informativ die generelle gute Passung der Item- und der Personenverteilung. Die Tafel führt allerdings auch vor, dass es eine Probandengruppe gab, die deutlich oberhalb der Aufgabenschwierigkeiten rangierte (ca. 67) und auch eine kleine Gruppe, die unterhalb des leichtesten Items situiert ist (ca. 23). Auf beide Ergebnisse ging das Forschungsteam nicht ein.

3 Differente Itemfunktion (DIF): Es ergab sich, dass 7 von 16 Items für Jungen schwerer als für Mädchen waren, man kann sagen, die Mädchen sind bei dieser Schreibaufgabe partiell leistungsfähiger. Es fehlen hier aber die Itemnummern.

4 Kompetenzstufen: Es hat sich die vermutete Zweiteilung der Fähigkeiten in etwa bewahrheitet. Allerdings taucht bei der Kompetenzzuweisung eine gewisse Überraschung auf: Eine der schwersten Aufgabe war nach den Resultaten, etwas über die eigenen Hobbies mitzuteilen, um dem Mädchen in Hawai, das dort sehr eigene Steckenpferde pflegt, eine gewisse Anregung zu geben: Der Logitwert zeigte 2.0. Die Gruppe beließ es aber bei ihrer Vormeinung, dass die Erwähnung der Hobbies zu den grundlegenden Fähigkeiten gehöre. Meistens korrigieren die Forschenden ihre Theorien nach den Maßgaben der Testung.

Trotz dieses letzten Einwandes besitzt das Projekt Mustergültigkeit und es ist auch deshalb anregend, weil es sowohl die Theorien, wie die Aufgabenliste veröffentlicht und zudem die statistischen Ergebnisse. Diesen Bericht über ein Forschungsprojekt kann man als Anregung auf allen Ebenen nehmen und so erhielt es u.a. auch positive Rückmeldungen von beteiligten Lehrkräften.

Der Rote Faden

Durch die gesamte Einführung zieht sich der Rote Faden, dass es die zitierten Empiriker/innen nicht für erwähnenswert halten, die Inhalte und Formulierungen der Aufgaben darzustellen, die sie

berechnet haben und schon gar nicht, die zugrundeliegende Theorien vorzustellen, die die Auswahlen steuern. So ist es ein großes Desiderat für die didaktische Zunft, ihrerseits die eigenen Modelle, die ja reichlich vorhanden sind, in den stabilisierenden Rahmen des Raschmodells einzufügen. Wie sagte schon Heinrich von Kleist (1810): "Man könnte die Menschen in zwei Klassen abteilen; in solche, die sich auf eine Metapher und 2) in solche, die sich auf eine Formel verstehn. Deren, die sich auf beides verstehn, sind zu wenige, sie machen keine Klasse aus." Diese Klasse müsste ja nur die Ergebnisse der Berechnungen anwenden!

Anmerkungen

1 Es sind mehrere anregende Arbeiten aus der Deutschdidaktik erschienen, die im Rahmen der IRT Latente Klassen (Winkler 2011; Magirius 2020) oder Zweidimensionalitäten untersuchen (Bachinger et al. 2020; Basendowski. 2020) Ihre Darstellung würde den hier gesetzten Rahmen überschreiten.

2 PISA hat in den letzten Durchgängen die unterste Stufe I noch einmal mit den Bezeichnungen Ia, Ib und Ic unterteilt, um Leistungen in dieser Region besser erfassen zu können, und umfasst auf diese Weise acht Stufen, s. Reiss u.a. (2019: 52-54).

3 Obwohl es sich um deutsche Lehrkräfte handelt, waren nur englische Darstellungen zu finden, in denen das Unterrichtsfach mit "science" benannt wurde. Da es sich explizit um Biologie-Lehrkräfte in der Ausbildung handelt, wurde dieses Fach so übersetzt.

Literaturangaben

Ayala, de, J.R. (2022[2]): The Theory and Practice of Item Response Theory. New York: Guilford Press

Bereiter, Carl (1980):Development in writing. Cognitive processes in writing: An interdisciplinary approach. L. Gregg, E.R. Steinberg Hillsdale, NJ, Erlbaum: 1-64

Bachinger, Antonia; Krelle, Michael; Engelbert-Kocher, Maria & von Eichhorn, Gabriele (2020): Zuhörkompetenzen messen. Ergebnisse der Bildungsstandard-Pilotierung in der 4. Schulstufe. Münster: Waxmann

Blatt, Inge/Ramge, Gesa/Voss, Andreas (2008): Modellierung und Messung der Textkompetenz im Rahmen einer Lernstandserhebung in Klasse 6. In: Didaktik Deutsch : Halbjahresschrift für die Didaktik der deutschen Sprache und Literatur 14 (2009) 26, S. 54-81

Boelmann, Jan M. (2021/22): Empirische Forschung in der Deutschdidaktik, Band 1-3. Baltmannsweiler: Hohengehren

Bond, Trevor G./Yan, Zi/Heene, Moritz (2021[5]): Applying the Rasch Model. Fundamental Measurements in the Human Sciences. New York: Routledge

Bond, Trevor G. (2020): Rasch Basics for the Novice. In: Khine, Myint Swe (Hrsg.)(2020): Rasch Measurement, 9-30

Boone, William J./Staver, John R./Yale, Melissa S. (2014): Rasch Analysis in the Human Sciences. Dordrecht u.a.: Springer

Canz, Thomas (2015): Validitätsaspekte bei der Messung von Schreibkompetenzen. Humboldt Universität Berlin: Dissertation

Fan, Jason/Bond, Trevor (2019): Unidimensionality and Local Independence.In Aryadoust & Raquel (Hrsg.): Quantitative Analysis for language assessment I. Fundamental Techniques. Abingdon und New York: Routledge, S. 102-121. (Daten: https://www.routledge.com/Quantitative-Data-Analysis-for-Language-Assessment-Volume-I-Fundamental/Aryadoust-Raquel/p/book/9781138733121/vol 1_Tutorials-06, (chapter 04))

Green, Rita (2013): Statistical Analyses for Language Testers. Basingstoke: Palgrave Macmillan

IQB: Institut zur Qualitätsentwicklung im Bildungswesen. https://www.iqb.hu-berlin.de/research

Kelava, Augustin/Moosbrugger, Helfried (2020): Einführung in die Item Response Theorie IRT. In: Moosbrugger, Helfried/Kelava, Augustin (Hrsg.)(2020³): Testtheorie und Fragebogenkonstruktion. Berlin: Springer, 369-409

Kleist, Heinrich von (1810) Fragmente. https://kleist-digital.de/berliner-abendblaetter/1810-61 [S. 242], Letzter Aufruf 30.12.2022

Klieme, Eckhard/Leutner, Detlev/Kenk, Martina (Hrsg.) (2010): Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. Zeitschrift für Pädagogik, 56. Beiheft

Leaper, D. (2010): Putting Students in their Place - with Rasch. In: Studies in Foreign Language Education, 24, 151-174

Linacre, John Michael(1998 ff): Winsteps - neueste Version 2022

Linacre, John Michael (2012). A user's guide to Winsteps Ministeps Rasch-model computer programs [version 3.74.0]. Retrieved from http://www.winsteps.com/winsteps.htm (stebi)

Linacre, John Michael (o.J.): First Step on Using Winsteps: youtube

Linacre, John Michael (2018) Analyzing Differential Item Functioning -DIF - with Winsteps. https://www.youtube.com/watch?v=YpM92j0Vq9s

Magirius, Marco (2020): Überzeugungen Deutschstudierender zum Interpretieren literarischer Texte: Eine Mixed-Methods-Studie. Berlin: J.B. Metzler

Raquel, Michelle (2019): The Rasch measurement approach to differential item functioning (DIF) analysis in language assessment. In: Aryadoust, Velid/Raquel, Michelle (Hrsg.)(2019): Quantitative Analysis for language assessment I. Fundamental Techniques. Abingdon und New York: Routledge, 103-131

Rasch, Georg (1960): Probabilistic models for some intelligence and attainment tests, Danish Institute for Educational Research: Copenhagen. expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press

Reiss, Kristina/Weis, Mirjam/Klieme, Eckhard/Köller Olaf (Hrsg.)(2019): PISA 2018. Grundbildung im internationalen Vergleich. Münster: Waxmann

Waechter, Friedrich Karl (1978): Wahrscheinlich guckt wieder kein Schwein. Zürich: Diogenes

Wieser, Dorothee (2019): Gegenwärtiger Stand der empirischen Unterrichtsforschung zum Literaturunterricht. In: Kämper-van den Boogaart/ Spinner, Kaspar H. (Hrsg.): Lese-und Literaturunterricht, Teil 2. Deutschunterricht in Theorie und Praxis, Band 11,2 (herausgegeben von Winfried Ulrich). Hohengehren: Schneider, 353-380

Willenberg, Heiner (2007): Der vergessene Wortschatz. In: Heiner Willenberg (Hrsg.): Kompetenzhandbuch für den Deutschunterricht, Hohengehren: Schneider, 148-156

Winkler, Iris (2011): Aufgabenpräferenzen für den Literaturunterricht: Eine Erhebung unter Deutschlehrkräften. Wiesbaden: VS Verlag für Sozialwissenschaften