Bildungsstandards M8 – Wie kommen die offiziellen Zahlen zustande und was sagen sie (nicht) aus?

ANDREAS VOHNS (UNIV. KLAGENFURT)

Die offiziellen Ergebnisrückmeldungen zur M8-Testung lassen viele Fragen offen: Ist es wirklich zu bedeutsamen Fortschritten gekommen? Sind Länderunterschiede so gravierend wie kolportiert? Kann „Statistische Darstellungen und Kenngrößen“ wirklich der „am besten“ absolvierte Inhaltsbereich[1] sein? Der Beitrag gewährt einen Blick auf das der Testung zu Grunde liegenden Mess- und Testverständnis / die eingesetzten Skalierungsverfahren, ohne welchen eine sinnvolle Interpretation der offiziellen Zahlen kaum möglich erscheint. Vor diesem Hintergrund werden offiziell verlautbarte Ergebnisse kritisch gewertet und es wird nach Antworten auf die o. g. Fragen gesucht.

1. Anlass und Fragestellungen

Im Dezember 2012 sind seitens des bifie die Ergebnisse der „Standardsüberprüfung 2012, Mathematik, 8. Schulstufe“ (kurz: BIST-M8) sowohl in Form von „individuellen“ Ergebnisrückmeldungen den am Test beteiligten Lernenden, Lehrenden und Schulleitungen als auch in Form von Bundes- und Länderergebnisberichten einer breiten Öffentlichkeit gegenüber kommuniziert worden. Die Ergebnisse haben weiters (für einen relativ kurzen Zeitraum) Eingang in die mediale Debatte gefunden, dort meist in Form von Länderrankings und Vergleichen zwischen AHS- und APS-Leistungen. Das bifie hat für die Kommunikation der Ergebnisse weitgehend auf eine in bestimmter Weise normierte Skala (vulgo: 500 ± 100-Skala) zurückgegriffen, die in ähnlicher Weise auch bei TIMSS, PISA und den Bildungsstandards-Überprüfungen in Deutschland eingesetzt wurde und wird. Zur Ermittlung der Punktwerte auf dieser Skala werden die tatsächlich erzielten individuellen Testleistungen („Rohscores“=Anzahlen korrekt gelöster Aufgaben) zunächst mit Hilfe eines bestimmten Modells aus der Familie der probabilistischen Testmodelle reduziert und transformiert, dem nach seinem Erfinder Georg Rasch (1901-1980) benannten *Rasch-Modell*.

Trotz seiner zunehmenden Verbreitung in der empirischen Bildungsforschung ist dieses Testmodell in seiner Funktionalität / Eignung für die Modellierung von (mathematischen) Schüler(innen)leistungsdaten nicht unumstritten[2]. In jedem Fall gilt es sich hinsichtlich der Interpretation der auf der 500 ± 100-Skala ermittelten Punktwerte, Gruppenunterschiede und Entwicklungen seit der „Ausgangsmessung für die Überprüfung der Bildungsstandards auf der 8. Schulstufe“ (kurz: Baseline-M8) im Jahr 2009 der spezifischen *Transformationen und Reduktionen* bewusst zu sein, denen die tatsächlich erzielten individuellen Testergebnisse unterworfen sind – wenn man die Ergebnisse nicht über- oder fehlinterpretieren will.

Der vorliegende Beitrag will hier zur Aufklärung beitragen, indem zunächst das klassische Rasch-Modell in Grundzügen dargestellt wird. Daran anschließend werden spezielle Aspekte der Anwendung des Modells im Rahmen von Baseline-M8 / BIST-M8 vorgestellt, insbesondere die Renormierung auf die 500 ± 100-Skala in der dort eingesetzten speziellen Variante, und es wird erläutert, wie es auf dieser Skala zur Einteilung von „Kompetenzsstufen“ gekommen ist.

Abschließend wird auf Basis eigener Analysen der veröffentlichten Daten und einer dem IDM Klagenfurt zur Verfügung gestellten Teilmenge nicht skalierter (allerdings bereits stark aggregierter) Lösungshäufigkeitsdaten folgenden Fragen nachgegangen:

- Haben sich die Schüler(innen)leistungen zwischen 2009 und 2012 tatsächlich „erheblich verbessert“ (Breit u. Schreiner 2012, S. 65), wie im Bundesergebnisbericht behauptet?

Für hilfreiche Anmerkungen zum Manuskript habe ich Thomas Jahnke und Joachim Wuttke zu danken.

[1] Zur Bedeutung der Begriffe „Inhaltsbereich“ und „Handlungsbereich“ (vgl. IDM 2007).

[2] In der deutschsprachigen Mathematikdidaktik wird vor allem die Ableitung von Kompetenzstufen in PISA (vgl. Meyerhöfer 2004, Lind u. a. 2005, Bender 2005), sowie allgemein der mit PISA verbundene Genauigkeitsanspruch kontrovers diskutiert (vgl. Wuttke 2007); in England galt dies bereits Anfang der 1980er Jahre allgemein für die Anwendung des Modells für Schüler(innen)leistungsdaten (vgl. Goldstein 1979, Preece 1980, Bryce 1981, Goldstein u. Blinkhorn 1982).

- Wie bedeutsam (überraschend) sind aus fachdidaktischer Sicht die mit großem medialen Interesse verbundenen Bundesländer-, Schulform- und Geschlechterunterschiede?
- Kann „Statistische Darstellungen und Kenngrößen“ wirklich der „am besten“ absolvierte Inhaltsbereich sein oder ist das Ergebnis ein Artefakt der Subskalenbildung?
- Wie ist insgesamt die Passung der statistischen Modellierung und der Form der Rückmeldung der Ergebnisse zu den Zielsetzungen der Standards M8 und dem dort verwendeten *theoretischen* Modell mathematischer Kompetenz zu beurteilen?

2. Grundzüge der Rasch-Modellierung

2.1. Kernidee: Zermelos abgebrochenes Schach-Turnier

Ich beginne zunächst mit einem Gedankenexperiment, das auf den ersten Blick etwas abseits unseres Themas liegt, allerdings eng mit einem historischen Vorläufer des Rasch-Modells in Verbindung steht[3]:

> Ein Schach-Verein mit 30 Mitgliedern will seine 10 spielstärksten Mitglieder zu einem internationalen Turnier reisen lassen. Seine Mitglieder unterteilen sich in die 10 Vorjahresteilnehmer des Turniers (Teilmenge I) und 20 weitere Mitglieder, die im letzten Jahr nicht an dem Turnier teilgenommen haben bzw. noch gar keine Vereinsmitglieder waren (Teilmenge P). Ein Vorauswahl-Turnier wird veranstaltet, bei dem zunächst jedes der Mitglieder aus I gegen jedes Mitglied aus P antritt, anschließend sollen auch alle Mitglieder aus I und P jeweils innerhalb ihrer Gruppen gegeneinander antreten. Aus irgendeinem Grund muss das Vorauswahl-Turnier aber abgebrochen werden, bevor die Mitglieder aus I und P jeweils innerhalb ihrer Gruppen konkurrieren konnten. Wie kann man nach diesem unvollständigen Turnier zu einem zuverlässigen Maß für die Spielstärke sämtlicher Mitglieder gelangen?

Wäre das Turnier zu Ende ausgetragen worden, so wäre die Lösung relativ naheliegend: Man würde schlicht die Anzahl der gewonnenen Spiele[4] als Maß der Spielstärke hernehmen. Da aber alle Personen aus P jeweils nur 10 Partien (gegen Gegner aus I), alle Personen aus I jeweils 20 Partien (gegen Gegner aus P) ausgetragen haben, kommt das nicht in Frage. Nun könnte man zum relativen Anteil der gewonnenen Spiele übergehen, dies würde aber unberücksichtigt lassen, dass die Personen aus P sich jeweils anderen Gegnern zu stellen hatten, als die Personen aus I. Wie kann es also gelingen, die Personen aus I und aus P auf einer gemeinsamen Skala der Spielstärke anzuordnen?

Ernst Zermelo (1929) widmete sich dieser Frage in seinem Aufsatz „Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung“. Zermelos Lösung liefe in unserem Beispiel vereinfacht auf das Folgende heraus: Personen von P gelten wie gewohnt als gleich spielstark, wenn sie dieselbe Anzahl von Spielen (gegen beliebige Personen aus I) gewonnen haben, ebenso gelten Personen aus I als gleich spielstark, wenn sie dieselbe Anzahl von Spielen (gegen Personen aus P) gewonnen haben. Eine Teilmenge gleich spielstarker Personen von P gilt nun genau dann als spielstärker als eine Person aus I, wenn der relative Anteil der Siege gegen diese Person in der Teilmenge größer als $0,5$ ist. Umgekehrt gilt eine Teilmenge von Personen gleicher Spielstärke von I als spielstärker als eine Person aus P, wenn der relative Anteil der Siege in dieser Teilmenge gegen diese Person größer als $0,5$ ist.

Um ein Maximumproblem handelt es sich dabei insofern, als die empirisch aufgetretenen relativen Gewinnhäufigkeiten in der Regel keine eindeutige Anordnung sämtlicher Personen aus I und G gemäß der oben angegebenen Prinzipien erlaubt und man daher mit Hilfe eines Schätzverfahrens eine *latente Variable* „Spielstärke“ als streng monotone Funktion der relativen Gewinnhäufigkeiten ermittelt, unter der die tatsächlich aufgetretenen Spielausgänge maximale Wahrscheinlichkeit haben. Man ersetzt dann die Gewinnhäufigkeiten durch (geglättete) Gewinnwahrscheinlichkeiten, die sich aus den Spielstärken

[3] Ich bezeichne dieses Gedankenexperiment hier in Anlehnung an Gallin u. Ruf als „Kernidee“ insofern es für mich persönlich der Schlüssel zum Verständnis des Rasch-Modells war.

[4] Remis werden hier vernachlässigt, obwohl Zermelo (1929) auch für solche Spiele eine Lösung parat hat.

berechnen lassen. Durch diesen Übergang zu (geglätteten) Gewinnwahrscheinlichkeiten statt ungeglätteten empirischen Gewinnhäufigkeiten ist stets eine eindeutige Anordnung möglich und die geschätzte Spielstärke ist gerade das geeignete Maß, demgemäß die Anordnung möglich ist (ein Zahlenbeispiel folgt im nächsten Abschnitt).

2.2. Das (eigentliche) Rasch-Modell

In etwa dasselbe passiert nun bei der Rasch-Modellierung von Leistungsdaten: Man ersetzt im obigen Gedankenexperiment die Spieler der Gruppe *I* durch Aufgaben („Items“) und die Spieler der Gruppe *P* durch Schüler(innen) („Persons“) sowie die Spielstärke im Falle der Aufgaben durch deren „Schwierigkeit“ (σ_i) und im Falle der Schüler(innen) durch deren „Fähigkeit“ (θ_v)[5], dann hat man das Grundprinzip des Rasch-Modells bereits im Kern erfasst. In dieser neuen „Einkleidung“ ist auch sofort einsichtig, warum das „Turnier“ abgebrochen werden muss: Eine Person kann an einer Aufgabe scheitern („Aufgabe gewinnt“) oder sie meistern („Person gewinnt“), aber weder Personen noch Aufgaben können direkt miteinander verglichen werden[6]. Für die weitere Erläuterung greife ich im Folgenden auf einen Modelldatensatz aus der 3. Pilottestung des Klagenfurter Projekts „Standardisierte schriftliche Reifeprüfung aus Mathematik (sRP-M)“ und eine auf Basis dieser Daten von mir erstellte Rasch-Modellierung[7] zurück.

Korrekte Aufgaben		**302**	**304**	**301**	**303**	**308**	**312**	**306**	**307**	**310**	**311**	**305**	**309**	**Personen**
Anzahl	**Anteil**													
12	1,00	1,00	1,00	1,00	1,00	1,00	1,00	**1,00**	1,00	1,00	1,00	1,00	1,00	23
11	0,92	0,83	0,92	0,88	0,92	0,96	0,96	**0,75**	0,92	0,96	0,92	1,00	1,00	24
10	0,83	0,73	0,77	0,83	0,87	0,83	0,83	**0,57**	0,97	0,93	0,73	0,93	1,00	30
9	0,75	0,68	0,52	0,68	0,64	0,88	0,64	**0,52**	0,84	0,76	0,88	0,96	1,00	25
8	0,67	0,61	0,65	0,52	0,65	0,57	0,52	**0,78**	0,78	0,52	0,61	0,87	0,91	23
7	0,58	0,42	0,63	0,46	0,54	0,54	0,67	**0,58**	0,46	0,50	0,58	0,75	0,88	24
6	0,50	0,25	0,33	0,58	0,29	0,29	0,46	**0,75**	0,42	0,67	0,67	0,50	0,79	24
5	0,42	0,16	0,32	0,16	0,26	0,32	0,37	**0,58**	0,37	0,47	0,63	0,74	0,63	19
4	0,33	0,40	0,30	0,30	0,30	0,10	0,30	**0,40**	0,10	0,20	0,80	0,30	0,50	10
3	0,25	0,00	0,13	0,00	0,50	0,38	0,25	**0,63**	0,13	0,00	0,13	0,38	0,50	8
2	0,17	0,00	0,00	0,22	0,11	0,11	0,22	**0,22**	0,11	0,22	0,33	0,00	0,44	9
1	0,08	0,00	0,00	0,00	0,00	0,33	0,00	**0,33**	0,00	0,00	0,00	0,00	0,33	3
0	0,00	0,00	0,00	0,00	0,00	0,00	0,00	**0,00**	0,00	0,00	0,00	0,00	0,00	1
Personen	**Anzahl**	119	129	131	135	138	140	**144**	144	146	157	169	189	
	Anteil	0,53	0,58	0,59	0,61	0,62	0,63	**0,65**	0,65	0,65	0,70	0,76	0,85	

Tab. 1: Item-Personengruppen-Matrix zum s-RP-M Pilottest 3 (empirische Werte)

In Tabelle 1 finden sich als Spalten die Aufgaben (nach Lösungshäufigkeit aufsteigend sortiert), als Zeilen finden sich jeweils Gruppen von Personen, die gleich viele Aufgaben korrekt gelöst haben (nach Anzahl korrekter Lösungen absteigend sortiert). In den Zellen finden sich jeweils die Lösungshäufigkeiten der Teilgruppen von Personen bei einer bestimmten Aufgabe. So lösen etwa 68% der Personen, die insgesamt 9 Aufgaben korrekt lösen, die Aufgabe 301 korrekt, welche insgesamt von 59% aller Personen korrekt gelöst wurde.

Während es unmittelbar einleuchtet, wie man Personen und Aufgaben untereinander in ihrer Leistung bzw. Schwierigkeit sortiert, ist nicht klar, wie man die Aufgaben zwischen die Personen mischen soll. Nach der im letzten Abschnitt angegebenen Regel müsste man für die Aufgabe 301 die Spalte von oben nach unten durchlaufen und schauen, wo die Lösungshäufigkeit der jeweiligen Teilgruppe von Personen von über 50% auf unter 50% wechselt. Für Aufgabe 301 ist dieser Wechsel aber nicht eindeutig festgelegt: Das erste Mal fällt der Wert für die Personengruppe mit 7 korrekten Antworten unter 50%, dann steigt er wieder und fällt für die Personengruppe mit 5 korrekten Antworten wieder deutlich unter 50%.

[5] „Schwierigkeit“ ist hier und im Folgenden strikt empirisch als geringe relative Lösungshäufigkeit einer Aufgabe, „Fähigkeit“ als hohe relative Lösungshäufigkeit einer Person zu lesen.

[6] Analog wird hier von Aufgaben ausgegangen, die nur als komplett richtig oder falsch gewertet werden, obwohl es auch für das Rasch-Modell Varianten für teilweise richtig zu wertende Aufgaben gibt.

[7] Vgl. Vohns (2012), die Analyse ist nicht Teil der offiziellen Auswertungen des Projekts sRP-M, die Aufgaben können unter http://www.aau.at/Zentralmatura-M abgerufen werden.

Wendet man das Rasch-Modell auf den Datensatz an, so entspricht dies, was den inneren Teil der Tabelle betrifft, im Wesentlichen einer Glättung der Daten. Zunächst werden dazu mit Hilfe eines Maximum-Likelihood-Verfahrens iterativ solche Werte für die θ_v und σ_i ermittelt, unter denen das tatsächliche Antwortverhalten der Testpersonen die höchste Wahrscheinlichkeit hat, wobei sich die bedingte Wahrscheinlichkeit, dass eine Person v der Fähigkeit θ_v eine Aufgabe i der Schwierigkeit σ_i korrekt löst, anschließend gemäß

$$P(X_{vi} = 1) = \frac{\exp(\theta_v - \sigma_i)}{1 + \exp(\theta_v - \sigma_i)}$$

zurückberechnen lassen soll[8]. Man kann also für jede Gruppe von Personen mit gleichem Fähigkeitswert θ_v die bedingte Lösungswahrscheinlichkeit für eine Aufgabe mit dem Schwierigkeitswert σ_i berechnen. Inhaltlich kann man diese Lösungswahrscheinlichkeit dann wieder als den (im Modell erwarteten) relativen Anteil der Personen gleicher Fähigkeit θ_v (gleicher Anzahl korrekter Aufgaben) interpretieren, der diese Aufgabe korrekt lösen sollte.

Korrekte Aufgaben			**302**	**304**	**301**	**303**	**308**	**312**	**306**	**307**	**310**	**311**	**305**	**309**	**Personen**
θ	**Anzahl**	**Anteil**													
3,36	12	1,00	0,94	0,95	0,95	0,96	0,96	0,96	**0,96**	0,96	0,97	0,97	0,98	0,99	23
2,49	11	0,92	0,86	0,89	0,89	0,90	0,91	0,91	**0,92**	0,92	0,92	0,94	0,96	0,98	24
1,69	10	0,83	0,74	0,78	0,79	0,80	0,82	0,82	**0,84**	0,84	0,84	0,88	0,91	0,95	30
1,16	9	0,75	0,62	0,68	0,69	0,71	0,72	0,73	**0,75**	0,75	0,76	0,81	0,86	0,92	25
0,74	8	0,67	0,52	0,58	0,59	0,61	0,63	0,64	**0,66**	0,66	0,68	0,73	0,80	0,89	23
0,37	7	0,58	0,43	0,49	0,50	0,52	0,54	0,55	**0,58**	0,58	0,59	0,66	0,73	0,84	24
0,01	6	0,50	0,35	0,40	0,41	0,43	0,45	0,46	**0,49**	0,49	0,50	0,57	0,65	0,79	24
-0,34	5	0,42	0,27	0,32	0,33	0,35	0,37	0,38	**0,40**	0,40	0,41	0,48	0,57	0,72	19
-0,72	4	0,33	0,20	0,24	0,25	0,27	0,28	0,29	**0,31**	0,31	0,33	0,39	0,47	0,64	10
-1,15	3	0,25	0,14	0,17	0,18	0,19	0,20	0,21	**0,23**	0,23	0,24	0,29	0,37	0,54	8
-1,69	2	0,17	0,09	0,11	0,11	0,12	0,13	0,14	**0,15**	0,15	0,15	0,20	0,25	0,40	9
-2,52	1	0,08	0,04	0,05	0,05	0,06	0,06	0,06	**0,07**	0,07	0,07	0,10	0,13	0,23	3
-3,41	0	0,00	0,02	0,02	0,02	0,02	0,03	0,03	**0,03**	0,03	0,03	0,04	0,06	0,11	1
Personen		**Anzahl***	120	129	131	135	138	139	**143**	143	145	155	167	186	
		Anteil*	0,54	0,58	0,59	0,60	0,62	0,63	**0,64**	0,64	0,65	0,70	0,75	0,83	
		σ	0,65	0,42	0,37	0,28	0,20	0,16	**0,06**	0,06	0,01	-0,28	-0,62	-1,30	

Tab. 2: Item-Personengruppen-Matrix zum s-RP-M Pilottest 3 (Rasch-modellierte Werte)

Es ergibt sich also eine neue Tabelle von (im Modell geschätzten) „Lösungswahrscheinlichkeiten“ (vgl. Tabelle 2[9]). Man erkennt sofort, dass diese neue Tabelle nun im Inneren in jeder Zeile von links nach rechts jeweils monoton wachsende und in jeder Spalte von oben nach unten monoton fallende Lösungswahrscheinlichkeiten enthält. Dadurch ist es möglich, eindeutig zu sagen, welche Personengruppen „stärker“ als Aufgabe 301 sind (diese also *eher* lösen werden, $P > 0,5$), welche „gleich stark“ (50 : 50-Lösungschance, $P = 0,5$) und welche „schwächer“ ($P < 0,5$): Für 7 richtige Aufgaben lösende Personen beträgt die Lösungswahrscheinlichkeit exakt 50%, für 6 Aufgaben lösende Personen ist sie kleiner. Aufgabe 301 landet also auf demselben Rangplatz wie die 7er-Gruppe, wer insgesamt mehr Aufgaben korrekt löst, löst Aufgabe 301 mit einer größeren Wahrscheinlichkeit als 50%, wer weniger Aufgaben korrekt löst, löst sie mit einer geringeren Wahrscheinlichkeit. In den Fähigkeits-/Schwierigkeitswerten drückt sich dies dadurch aus, das für Aufgabe 301 und die 7er-Gruppe $\theta_{\Sigma=7} = \sigma_{I301} = 0,37$ gilt. Die Möglichkeit einer eindeutigen Anordnung hat man sich aber dadurch erkauft, dass man tatsächlich aufgetretene empirische Lösungshäufigkeiten durch solche Werte ersetzt hat, die bei Passung des Modells eigentlich hätten herauskommen sollen.

Die hier vorgestellte Tabellendarstellung stellt eine gewisse Vereinfachung dar. Normalerweise beginnt man mit einer Tabelle, bei der die Zeilen einzelne Personen sind („Item-Personen-Matrix“). Die Zusam-

[8] Die genaue Funktionsweise der dazu einsetzbaren Schätzverfahren kann hier nicht näher erläutert werden, ist für die folgende Argumentation aber auch nicht entscheidend. Eine theoretisch orientierte Einführung gibt Rost (1996), eine praktische anhand der Softwarepakete eRm/R u. a. Strobl (2010).

[9] Ein einfacher qualitativer Test der Passung des Modells zu den Daten besteht daher darin, das Innere beider Tabellen miteinander zu vergleichen: Dort wo die Übereinstimmung groß ist, passt das Modell gut. Dort wo beide Tabellen sich stark unterscheiden, sind potentiell Abweichungen vom Modell zu suchen

menfassung zu Personengruppen ist insofern sinnfällig, als im klassischen Rasch-Modell (für den Fall, dass alle Personen sämtliche Aufgaben eines Tests bearbeitet haben) der für zwei Personen mit derselben Lösungshäufigkeit ermittelte Fähigkeitswert θ *stets übereinstimmen muss*. Dasselbe gilt analog auch für die Schwierigkeitswerte von zwei Aufgaben mit derselben Lösungshäufigkeit. Relative Lösungshäufigkeiten und ermittelte Fähigkeits- bzw. Schwierigkeitswerte hängen funktional zusammen, genauer: Die θ_v und σ_i sind Funktionswerte einer streng monoton wachsenden (θ_i) bzw. fallenden (σ_i) Funktion der jeweiligen relativen Lösungshäufigkeiten[10]. Auf einzelne Personen bzw. Aufgaben zurückübertragen bedeutet dies:

1. Für die Einschätzung der Fähigkeit einer Person ist nur entscheidend, *wie viele*, aber *nicht welche* Aufgaben sie korrekt gelöst hat.
2. Für die Einschätzung der Schwierigkeit einer Aufgabe ist nur entscheidend, *wie viele*, aber *nicht welche* Personen sie korrekt gelöst haben.

Die Glättung der Tabelle 2 kommt dabei dadurch zustande, dass man einerseits unterstellt, dass z. B. die 7 „schwierigsten“ Aufgaben typischerweise von den 7 „fähigsten“ Personengruppen mit den höchsten Wahrscheinlichkeiten gelöst werden. Andererseits unterstellt man, dass Personen, die z. B. 7 Aufgaben korrekt lösen, typischerweise ähnliche Aufgaben mit hoher Wahrscheinlichkeit lösen, wobei die 7 „leichtesten“ Aufgaben die höchste Wahrscheinlichkeit haben, zu diesen 7 Aufgaben zu gehören[11]. Ein extremer Grenzfall wäre es, wenn alle 7 Aufgaben lösenden Personen genau dieselben, nämlich die 7 insgesamt am häufigsten gelösten Aufgaben korrekt gelöst hätten (bzw. wieder umgekehrt: wenn die 7 schwierigsten Aufgaben *genau* von den 7 fähigsten Personengruppen überhaupt korrekt gelöst würden, von allen anderen Personengruppen nicht). Dann hätte man in der Tabelle eine obere Dreiecksmatrix mit lauter Einsern. Im Rasch-Modell wird ein solcher deterministischer Zusammenhang zwischen Personenfähigkeit und Aufgabenschwierigkeit nicht unterstellt, sondern nur ein probabilistischer, bei dem zumindest monoton wachsende Lösungswahrscheinlichkeiten bei wachsenden Fähigkeitswerten ewartet werden.

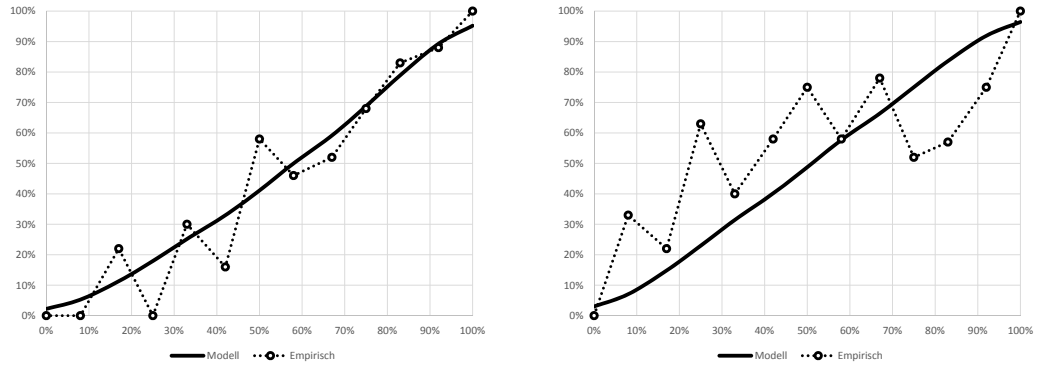

Abb. 1: Itemcharakteristiken Aufgabe 301 (links) und 306 (rechts), s-RP-M Pilotest 3

Während die im Modell angenommene Lösungswahrscheinlichkeitsreihenfolge der Aufgaben in der Gesamtgruppe immer die empirische Häufigkeitsreihenfolge abbilden muss[12], kann man gleichwohl selbst bei „Passung“ des Modells nicht erwarten, dass für eine einzelne Aufgabe (eine Spalte im Inneren der Tabelle) die empirischen Lösungshäufigkeiten von einer Teilgruppe gleich viele Aufgaben korrekt lösender Personen zur nächsten Teilgruppe gleich viele Aufgaben korrekt lösender Personen immer dem montonen Verlauf der geglätteten Lösungswahrscheinlichkeitsreihenfolge folgt – dann wäre ja auch keine Glättung nötig (vgl. etwa die Itemcharakteristiken für Aufgabe 301 und 306 in Abb. 1). Es wird im Modell

[10] Die zugehörigen Funktionsvorschriften sind aber zunächst unbekannt, die Funktionswerte müssen iterativ mit einem Maximum-Likelihood-Verfahren ermittelt werden.

[11] Die Wahrscheinlichkeit, dass unter den sieben gelösten Aufgaben einer einzelnen Personen dieser Teilgruppe wenigstens drei Aufgaben enthalten sind, die nicht zu den am häufigsten gelösten Aufgaben gehören, beträgt immerhin noch fast 50%.

[12] Die unteren Randzeilen Anteil / Anzahl in Tabelle 1 bzw. 2 stimmen numerisch fast überein, jedenfalls bleibt die Reihenfolge stets erhalten.

auch nicht angenommen, dass sich für eine Einzelperson stets eine Aufgabe findet, ab der sie dann sicher genau alle leichteren (also im Durchschnitt öfter gelösten) Aufgaben korrekt löst. Man geht aber davon aus, dass stark abweichende Lösungsmuster (als Extremfall: jemand löst ausgerechnet die schwierigsten Aufgaben, die leichtesten hingegen nicht) bei so wenigen Personen auftreten, dass diese „Pattern“ als „Rauschen“ klassifiziert werden können und keine relevante Information über die Fähigkeitsausprägung beinhalten. Statistisch lassen sich gewisse Kriterien angeben, wann abweichende Lösungsmuster in einem derart großen Umfang auftreten, dass man nicht mehr von zufälligen Abweichungen ausgehen sollte und das Rasch-Modell den Datensatz nicht mehr angemessen modelliert. Im Fall des 3. Pilotests s-RP-M ist dies z. B. für die oben dargestellte Aufgabe 306 der Fall: die empirisch aufgetretenen relativen Häufigkeiten stimmen hier derart gering mit den geschätzten Lösungswahrscheinlichkeiten überein, dass nicht mehr von bloß zufälligen Abweichungen ausgegangen werden kann.

Dabei ist es wichtig sich zu verdeutlichen, dass dies *keine* Eigenschaft der Aufgabe 306 an sich ist, sondern eine Eigenschaft der Aufgabe 306 *relativ* zu den anderen elf Aufgaben des Tests und *relativ* zur Gruppe der Getesteten. Denn: Die Logik, dass wenige Aufgaben korrekt lösende Personen im Wesentlichen ähnliche Aufgaben korrekt lösen, hängt an der Voraussetzung der *Homogenität* der Aufgaben bzw. der Eindimensionalität der durch sie gebildetenen Skala *und* ebenso an der *Homogenität* des Lösungsverhaltens / der Fähigkeitsstruktur der Getesteten. Als Gegenbeispiel wieder ein Gedankenexperiment: Würde man bewusst einen Test zusammenstellen, bei dem die erste Hälfte der Fragestellungen sich auf sportliche Ereignisse der letzten 10 Jahre beziehen und die zweite Hälfte aus Fragen zu populären Filmen desselben Zeitraums, so könnte es Personen geben, die sich nur für einen der Bereiche stark, für den jeweils anderen überhaupt nicht interessieren und solche, die sich für beide Bereiche eher mittelmäßig interessieren. Alle drei Gruppen würden dann in Summe vermutlich ähnliche, nämlich eher mittelmäßige Ergebnisse erzielen, aber jeweils andere Fragen wären für diese Gruppen „einfach“ und würden daher häufig korrekt gelöst. Die Gesamtheit der Aufgaben kann dann nicht sinnvoll eindimensional skaliert werden. Ob eine Aufgabe eine „hohe psychometrische Qualität“ hat oder als „Rasch-konform“ gilt, kann man immer nur *relativ* zur Auswahl der übrigen Aufgaben eines Tests entscheiden und immer erst nachdem ein (Pilot-)Test durchgeführt wurde. Deswegen werden nach dem Rasch-Modell zusammengestellte Tests in aller Regel auch pilotiert und sich als nicht Rasch-konform herausstellende Aufgaben werden ggf. aus dem Aufgabenset ausgeschlossen, bevor dieses im Haupttest eingesetzt wird. An dieser Stelle entstehen allerdings Brüche zwischen Testmodell und fachdidaktischem Kompetenzmodell, auf die später noch einzugehen sein wird.

Zuvor stellt sich die Frage, warum man überhaupt Rasch-Werte berechnet, wenn diese doch zu genau derselben Anordnung von Aufgaben bzw. Personen führen, wie die relativen Lösungshäufigkeiten. Rost (1996, S. 126f.) sieht die wesentliche Aufgabe des Rasch-Modells darin, zu überprüfen, *ob* ein vorgegebenes Set von Aufgaben tatsächlich (hinreichend) homogen ist, also Testleistungen sinnvoll auf einer eindimensionalen Skala abgebildet werden können – nur dann könne man die Summe korrekter Lösungen (im strengen Sinne) als Messwert für die Fähigkeitsausprägung einer Personen auffassen[13]. In diesem Sinne der *Überprüfung der Eindimensionalität eines Fähigkeitskonstrukts* ist mir allerdings keine Anwendung des Rasch-Modells innerhalb der Mathematikdidaktik bekannt – vermutlich weil ohnehin (fast) niemand „mathematische Leistung“ ernsthaft für ein eindimensionales Konstrukt hält, empirisch erweist sich dies schon am Ende der Volksschule als kaum gerechtfertigt (vgl. Ratzka (2004), S. 149).

Für die Testung von Schüler(innen)leistungen wird das Rasch-Modell im Zuge der Pilotierung von Testaufgaben realiter in der Mathematikdidaktik eher im Sinne einer *pragmatisch motivierten Skalierungsnorm* eingesetzt, um die Eindimensionalität von Testaufgabensätzen durch gezielte Aufgabenausschlüsse erst *herzustellen*[14]. Für den Einsatz des Rasch-Modells sprechen dabei weniger mathematikdidaktisch-inhaltliche, als vielmehr pragmatische Gründe:

1. Personen und Aufgaben landen wie oben dargestellt auf einer gemeinsamen Skala. Man kann also Personen identifizieren, die bestimmte Gruppen von Aufgaben mit einer Mindestwahrscheinlichkeit

[13] Zur Kritik dieser Logik vgl. Goldstein 1980, S. 211.

[14] Zur Kritik dieses „Skalierungspragmatismus“ vgl. Vohns (2012), S. 341ff.

lösen. Das kann von Interesse sein, wenn die Aufgaben selbst nicht veröffentlicht, sondern Anforderungsumschreibungen der Aufgabengruppen zur Beschreibung der Fähigkeiten von Personen genutzt werden sollen (vgl. Abs. 3.3).

2. Ein breites Spektrum/eine große Anzahl an Aufgaben soll getestet werden, bei dem es aus Zeitgründen schlicht nicht möglich ist, allen Testteilnehmer(inne)n sämtliche Aufgaben vorzulegen. Durch das Rasch-Modell können die Fähigkeiten verschiedener Personen auch dann noch miteinander verglichen werden, wenn nicht alle Personen exakt dieselben Aufgaben bearbeitet haben, sondern es eine hinreichend große Teilmenge an Aufgaben gibt, die alle Personen bearbeitet haben. Man erstellt dafür verschiedene Testhefte, die jeweils Überschneidungen in den enthaltenen Aufgaben haben. In den Tabellen von oben hat man dann fehlende Werte /leere Zellen (sog. „missing-data-by-design"), kann aber totzdem noch θ_v und σ_i schätzen und hat bei hinreichender Passung des Rasch-Modells weitgehend unverzerrte Gruppenmittelwerte.

3. Es sollen in gewissen zeitlichen Abständen Testwiederholungen stattfinden, bei denen analog zu 2. nicht immer genau dieselben Aufgaben eingesetzt werden können oder sollen[15].

3. Anwendung und Modifikationen des Rasch-Modells in der Skalierung von Baseline-M8 und BIST-M8

Wer sich mit der speziellen im Falle der Standard-Testungen in Österreich eingesetzten Variante der Rasch-Modellierung und Normierung auf die 500 ± 100-Skala beschäftigen will, stößt schnell an Grenzen, da der technische Bericht von Baseline-M8 bislang nur unvollständig (vgl. Breit u. Schreiner 2010) und ein technischer Bericht für BIST-M8 überhaupt nicht vorliegt. Die im Folgenden gemachten Aussagen beziehen sich daher z. T. auf Informationen, die sich einerseits aus Mailwechseln des IDM-Klagenfurt mit dem bifie, andererseits aus einem Gesprächstermin im bifie im Dezember 2013 speisen. Im Sinne der wissenschaftlichen Kontrolle der Standard-Testungen ist sich der Einschätzung von Altrichter und Neuwirth anzuschließen, dass hier deutlich mehr Transparenz in der Dokumentation und eine volle Offenlegung anonymisierter (!) Rohdaten zu wünschen wäre, ohne die ernsthafte Sekundäranalysen unmöglich sind (vgl. Nimmervoll 2014a u. 2014b).

3.1. Übersicht: Umfang, Zweck und Logik der einzelnen Testdurchäufe

Zeitraum	Test	Anzahl getesteter Personen
2005, 2006, 2007	Pilotierung Baseline-M8	$3.000 - 5.000$
2009	**Baseline-M8**	10.000
2010, 2011	Pilotierung BIST-M8	mehrere Tausend
2012	**BIST-M8**	80.000 (Vollerhebung)

Tab. 3: Testungen im Umfeld der Standards

Im Umfeld der Standards-M8 haben bislang die in Tabelle 3 angegebenen Testungen stattgefunden. Wobei die jeweils nicht fett-gedruckten Pilot-Testungen einzig dem Zweck dienen, *Aufgabensets an Schüler(innen)* zu testen, also ggf. zu leichte, zu schwierige oder nicht modellkonforme Aufgaben auszusortieren. Über die Anzahl der insgesamt pilotierten Aufgaben sind bislang keine Angaben veröffentlicht worden, ebenso wenig zu den bei der Pilotierung angewandten Normen, Regeln oder Kennwerten, auf deren Basis Aufgaben beibehalten oder verworfen wurden.

Aus unserem Gespräch mit dem bifie ist mir bekannt, dass die Herstellung eines homogenen Aufgabensets dem bifie nach eigenen Angaben keine allzu großen Probleme bereitet hat, man allerdings bei den Aufgabenausschlüssen auch im Zweifel eher der Repräsentation der Breite der durch die Standards abgedeckten mathematischen Fähigkeiten als der guten Passung der Aufgaben zum Modell den Vorzug

[15] Die Aussage, dass zwei gleich viele Aufgaben lösende Personen denselben Fähigkeitswert zugewiesen bekommen, gilt für die Fälle 2. und 3. nur mehr für Personen, die dasselbe „Testheft" (mit denselben Aufgaben) erhalten haben.

gegeben hat, sich also vermutlich auch eher schlecht passende Aufgaben im Set befinden. Bestätigt wurde uns auch, dass die Zusammenstellung der Aufgabensets für BIST-M8 und Baseline-M8 gemäß üblichen testpragmatischen Überlegungen so angelegt wurde, dass man eine mittlere Lösungswahrscheinlichkeit um 50% erwarten konnte und eine „hinreichende“ Streuung der Aufgabenschwierigkeiten (ca. zwischen 5% und 95% Lösungshäufigkeit) angepeilt wurde. Inwiefern solche Setzungen zum offiziell verordneten Zweck der Standards-M8 passen, ein Minimalniveau an Fähigkeiten sicherzustellen, welches in der Regel zu erreichen ist, ist durchaus hinterfragenswert.

3.2. Normierung der Skalen: Woher die 500 ± 100-Werte kommen

Eine Umrechnung der rohen Rasch-Werte θ_i und σ_v erfolgte für die Ausgangsmessung Baseline-M8 2009 so, dass der Mittelwert der Personenfähigkeiten $\overline{\theta_i}$ für die Gesamtskala (ebenso wie auf allen acht Subskalen für Inhalts- und Handlungsbereiche getrennt) auf 500 fixiert wurde, die Standardabweichung jeweils auf ± 100 – die Werte wurden also schlicht noch einmal linear transformiert (allerdings mit unterschiedlichen Parametern für die Gesamtskala und jede der acht Subskalen). Dass die Testergebnisse von Baseline-M8 *normalverteilt* sind, ist *nicht* ursächlich dieser Transformation zuzuschreiben, sondern der Tatsache, dass die Schätzung der Schwierigkeits- und Fähigkeitsparameter gemäß einem speziellen Schätzverfahren erfolgt ist, welches eine Normalverteilung der Populationsparameter als Annahme *voraussetzt* (dasselbe gilt für BIST-M8). Die Werte der Aufgabenschwierigkeiten wurden ebenfalls auf die 500 ± 100-Skala transformiert, allerdings (analog zu PISA) gegen die Fähigkeitsskala soweit verschoben, dass eine Person mit einem Punktwert von 500 für eine Aufgabe mit dem Punktwert von 500 eine Lösungswahrscheinlichkeit von $62,5\%$ hat. Anders als bei PISA wurden seitens des bifie die Punktwerte von keiner einzigen Aufgabe veröffentlicht[16].

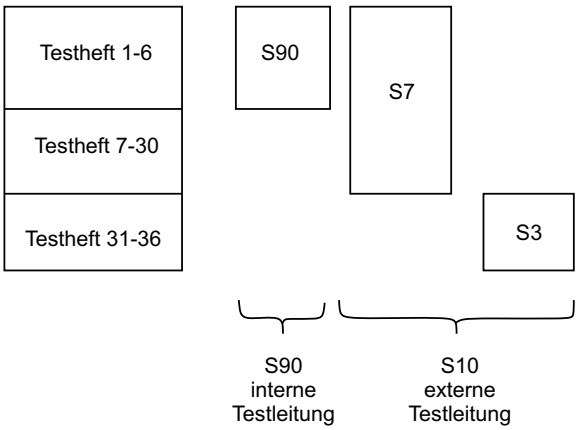

Abb. 2: Teilstudien BIST-M8 (Quelle: Email Simone Breit, Juli 2013)

Weder in Baseline-M8 noch in BIST-M8 bearbeiten alle Schüler(innen) sämtliche eingesetzten Aufgaben. Bei BIST-M8 bearbeiten etwa diejenigen 90% der Schüler(innen), bei denen der Test mit einer internen Testleitung ablief (Gruppe S90 in Abb. 2, vgl. auch den Beitrag von Edith Schneider in diesem Band), von den dort insgesamt eingesetzten 72 Aufgaben jeweils 48. Die eingesetzten Testhefte $1-6$ sind so zusammengestellt, dass einzelne Aufgaben vollständig durch die Testhefte rotieren, was dazu führt, dass von den 72 Aufgaben jede Aufgabe von etwa 50.000 Schüler(inne)n bearbeitet wurde. Von diesen 72 Aufgaben aus BIST-M8 wurden zudem 40 Aufgaben bereits in der Ausgangsmessung Baseline-M8 eingesetzt. Zusätzlich gab es eine Teilpopulation von 3% in BIST-M8 (Gruppe S3 in Abb. 2), denen die unveränderten Testhefte $31-36$ aus Baseline-M8 vorgelegt wurden. In den Testheften $7-30$, die den übrigen 7% in BIST-M8 Getesteten (Gruppe S7 in Abb. 2) unter externer Testleitung zusätzlich zu den Testheften $1-6$ vorgelegt wurden, finden sich weitere 77 Aufgaben, die sämtlich nicht in Baseline-M8 eingesetzt wurden.

[16] Dem IDM Klagenfurt liegen zumindest Punktwerte für die zehn veröffentlichten Aufgaben vor. Kennt man (wie wir) außerdem die Lösungshäufigkeiten sämtlicher Aufgaben, so lassen sich die Punktwerte für die übrigen Aufgaben in guter Näherung interpolieren. Die Geheimhaltung der Punktwerte wird damit zur völlig unnötigen Schikane.

Der Grund für diese zusätzlichen Aufgaben liegt wohl darin, dass alle 72 Aufgaben aus den Testheften $1-6$, welche bei interner Testleitung eingesetzt wurden, aufgrund mangelnder Kontrollierbarkeit der Geheimhaltung nach mündlicher Auskunft des bifie für weitere Testdurchläufe nicht mehr eingesetzt werden sollen. Testleistungen späterer Durchläufe wären aber nicht mehr mit denen von Baseline-M8 und BIST-M8 vergleichbar, wenn man 2015 zu völlig anderen Aufgaben überginge, daher setzt man in Testheft $7-30$ zusätzliche Aufgaben ein, die man dann 2015 wieder verwenden kann.

Die wechselseitige Verknüpfung der Testhefte durch gemeinsame Aufgaben innerhalb von BIST-M8 und mit Baseline-M8 erlaubt es nun umgekehrt (Passung des Rasch-Modells hier und im Folgenden vorausgesetzt) hochzurechnen, wie gut die 2012 getesteten Schüler(innen) abgeschnitten hätten, wenn man ihnen den Test Baseline-M8 vorgelegt hätte. Der für BIST-M8 ermittelte Mittelwert von 535 sagt dabei im Prinzip aus: Skaliert man beide Datensätze gemäß dem Rasch-Modell und normiert den ersten Test auf 500 ± 100, so ist das unter Voraussetzung des Rasch-Modells wahrscheinlichste Ergebnis, welches die Gesamtpopulation von 2012 im Test von 2009 erzielen würde, um 35 Punkte höher als das Testergebnis der Schüler(innen), die den Test 2009 tatsächlich bearbeitet haben. Einfacher ausgedrückt: Die Ergebnisse von 2012 lassen erwarten, dass die dort getesteten Schüler(innen), wenn man ihnen den Test von 2009 vorgelegt hätte, im Durchschnitt eine gewisse Anzahl von Aufgaben mehr korrekt gelöst hätten. Den offiziellen Angaben lässt sich aber nicht entnehmen, wie viele Aufgaben das sind. Durch Rückrechnung kann man in Kenntnis der Lösungshäufigkeitsdaten approximativ ermitteln: auf 48 Aufgaben eines idealtypischen Testhefts[17] umgelegt entsprechen 35 Punkte mehr etwa einer Größenordnung von 3 zusätzlich korrekt gelösten Aufgaben.

3.3. Standard-Setting: Woher die Kompetenzstufen kommen (und was sie aussagen)

Stufe	**Punktskala**	**Korrekte Lösungen**	
	(500 ± 100)	**Anteil**	**Anzahl**
3: übertroffen	≥ 691	$100-76\%$	$48-37$
2: erreicht	$690-518$	$75-45\%$	$36-22$
1: teilweise erreicht	$517-440$	$44-30\%$	$21-15$
0: nicht erreicht	≤ 439	$30-0\%$	$14-0$

Tab. 4: Punktwerte und Rohscores eines idealtypischen Testhefts (eig. Berechnungen)

In Tabelle 4 können die durch Rückrechnung bestimmten ungefähren Rohscores in einem solchen Testheft ermittelt werden, die jeweils an den Grenzen der sog. „Kompetenzstufen“ in etwa zu erzielen waren. Ein verbreitetes Missverständnis bei der Anwendung des Rasch-Modells besteht darin zu glauben, aufgrund seiner Anwendung könnten aus Lösungshäufigkeitsdaten per se kriteriale Normen („Bildungsstandards sind erfüllt“) abgeleitet werden (vgl. Rost 2004, S. 663). Wenn überhaupt ist dies im Falle von BIST-M8 nur insofern möglich, als man sich der Logik des folgenden Verfahrens anschließen mag:

> Dreiundzwanzig „Expert(inn)en“[18] wird jeweils ein Heft mit 80 Aufgaben einer bereits durchgeführten Testung vorgelegt. Die Aufgaben wurden dabei nach Lösungshäufigkeit absteigend sortiert („Ordered-Item-Booklet“). An jede Aufgabe notieren sie $1, 2$ oder 3. Aufgaben, die sie mit 1 markieren, sollten auch solche Personen lösen können, bei denen sie die Standards nur für teilweise erreicht halten, die mit 2 markierten Aufgaben sollten alle Schüler(innen) lösen können, die die Standards erreichen und bei den mit 3 markierten Aufgaben würde man auch von die Standards erreichenden Schüler(inne)n nicht unbedingt eine korrekte Lösung erwarten. Im Idealfall stimmen die Urteile aller „Expert(inn)en“ für jede Aufgabe überein und die Bewertungen ergeben eine monoton wachsende Folge von Zahlen. Man legt dann fest: Wer für weniger als die Hälfte der mit 1 bewerteten Aufgaben jeweils eine Lösungswahrscheinlichkeit von mindestens 67% hat, hat die Standards nicht erreicht. Analog bestimmt man eine Mindestanzahl von Aufgaben, die man von den mit 2 bewerteten (und allen leichteren) zusätzlich mit dieser Wahrscheinlichkeit korrekt lösen sollte, um

[17] Es wird angenommen, alle Testhefte wären jeweils gleich schwierig (gleiche mittlere Lösungshäufigkeit).

[18] „Vertreter der Fachdidaktik (30,43%), der Elternvertretungen (13,04%), verschiedener Abnehmergruppen wie zum Beispiel Berufschulen (17,39%), des Ministeriums (8,70%), praktizierende Lehrer/innen für M4 und M8 (13,04%), des Schulrats (8,70%) und der Psychometrie (8,70%)“ (Freunberger 2013, S. 3)

die Standards erreicht zu haben, schließlich zusätzlich noch eine Mindestanzahl von den mit 3 bewerteten (und allen leichteren) mit dieser Wahrscheinlichkeit korrekt zu lösenden Aufgaben, damit die Standards als „übertroffen“ gelten.

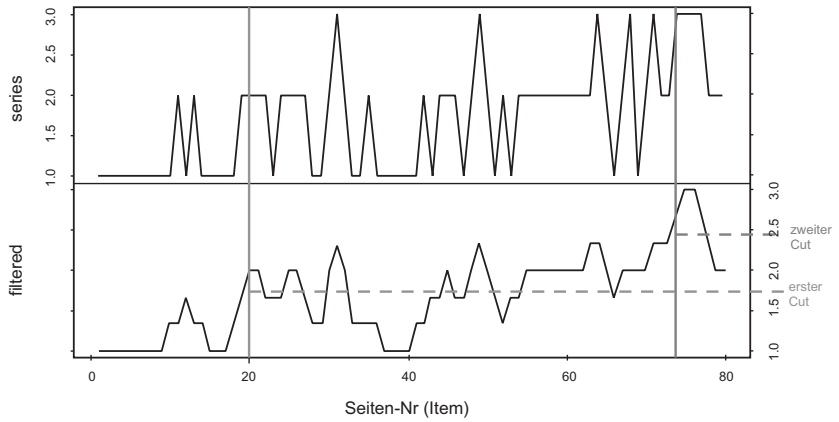

Abb. 3: Individuelle Bewertungsserie (ungeglättet, geglättet) (Freunberger 2013, S. 9)

Wem dieses Verfahren problematisch erscheint, dem sei zusätzlich noch gesagt, dass in der Realität keine der Expert(inn)enbewertungen eine streng monoton wachsende Folge von Zahlen ergeben hat (eine individuelle Bewertungsserie findet sich in Abb. 3, oberer Graph) und sich die Expert(inn)en untereinander nicht einig waren. Unter Vernachlässigung des ordinalen Niveaus der Expert(inn)enbewertungen werden daher zunächst gleitende Durchschnitte über mehrere Aufgaben hinweg („filtered scores“, vgl. Abb. 3, unterer Graph) und anschließend Mittelwerte über die so „gefilterten“ Expert(inn)enurteile gebildet (vgl. a. a. O., S. 7-9), schließlich werden die „ Schwellenwerte [...] auf $1,7$ für den ersten Cut[19] und auf $2,4$ für den zweiten Cut[20] gesetzt“ (a. a. O., S. 7). Man sieht dann im Testheft nach, welche Seite des Booklets an dieser Stelle läge und nimmt deren Wert auf der 500 ± 100-Skala als Cut-Off-Score her. Abschließend einigt man sich noch darauf, dass die Standards als nicht erfüllt gelten, wenn eine Person für weniger als 7 von 12 Aufgaben unterhalb der $1,7$ eine Lösungswahrscheinlichkeit von 67% und mehr hat (vgl. a. a. O., S. 11), man bestimmt wieder das zugehörige Item im Booklet und liest dessen Punktwert als Grenzwert ab. Gemäß Freunberger (2013) handelt es sich bei dem angewandten Verfahren um eine Modifikation eines in der Psychometrie etablierten Verfahrens („item-descriptor-matching“). Das heilt wohl kaum die eher nichtssagenden inhaltlichen Beschreibungen der einzelnen Stufen[21], noch immunisiert es gegen die von Meyerhöfer (2004) grundsätzlich gegen dieses Konstrukt vorgebrachte Kritik. Im Zweifelsfall sagen die Kompetenzstufen nichts anderes aus als das, was in Tabelle 4 in der 3./4. Spalte vermerkt ist.

4. Analyseteil

4.1. Haben sich die Schüler(innen)leistungen seit 2009 „erheblich verbessert“?

Grundlage dieses Urteils stellt offenbar eine (nicht näher erläuterte) Beurteilung des Mittelwertunterschieds zwischen 2009 und 2012 dar. Für solche Mittelwertunterschiede kann zunächst nach statistischer Signifikanz gefragt werden, die aber nur klärt, ob man Gruppenunterschiede als zufälliges Ergebnis ablehnen muss. Daraus ist noch kein Urteil über die inhaltliche Relevanz des Unterschieds abzuleiten: Bei hinreichend großen Stichproben werden auch sehr kleine Unterschiede als überzufällig qualifiziert. Ein in der Psychometrie übliches weiteres Maß ist die standardisierte *Effektgröße* bzw. *-stärke*, meist angegeben als *Cohen's d*. Es handelt sich dabei um die Mittelwertdifferenz relativ zur Streuung gemessen in % der gepoolten Standardabweichung. Im Falle der Standardtestungen ergibt sich (unter Beachtung

[19] Grenze für „Standards erreicht“

[20] Grenze für „Standards übertroffen“

[21] Exemplarisch „teilweise erreicht: Die Schüler/innen verfügen über grundlegende Kenntnisse und Fertigkeiten in allen Teilbereichen des Lehrplans Mathematik und können damit reproduktive Anforderungen bewältigen und Routineverfahren durchführen.“ (a. a. O., S. 3)

der unterschiedlichen Gruppengrößen, vgl. Tabelle 5) $d_{\text{cohen}} = 0,345$. Cohen (1988) selbst stuft einen solchen Unterschied als „small effect“ ein, gemäß Hattie (2009, S.97) würde man sehr knapp die „zone of desired effects“ erreichen. Gemäß Cohen hätte das bifie also unrecht, gemäß Hattie könnte man ihm vielleicht recht geben.

	Baseline-M8 (2009)	BIST-M8 (2012)
Mittelwert	500	535
Standardabweichung	100	101,5
Gruppengröße (N)	10.082	79.682
Effektstärke (d_{cohen})	**0,345**	

Tab. 5: Effektstärke Entwicklung 2009 – 2012 (Breit u. Schreiner 2012, eig. Berechnungen)

Effektgrößen können nicht nur in standardisierten Maßen betrachtet werden, dort wo Rohdaten leicht interpretierbar erscheinen, stellen diese eine bereichernde Alternative dar. Der hier vorliegende Mittelwertunterschied lässt sich (bei hinreichender Passung des Modells) näherungsweise wieder in einen Prozentsatz zusätzlich korrekter Lösungen (bzw. in Anzahlen zusätzlicher korrekter Aufgaben) zurückumrechnen, die in BIST-M8 getestete Personen in Baseline-M8 gegenüber den seinerzeit getesteten Personen korrekt gelöst hätten. Der Einfachheit halber unterstelle ich, alle Testhefte wären exakt gleich schwierig (betrachte ein idealtypisches Testheft). In so einem Testheft hätte eine in Baseline-M8 getestete Person im Durchschnitt etwa 20 von 48 Aufgaben (41%) korrekt gelöst, eine in BIST-M8 getestete Person hätte ca. 23 Aufgaben (48%) korrekt. Ein solcher Unterschied ist vermutlich zu groß, um als rein zufällig oder reines Artefakt des Skalierungsmodells gelten zu können.

Inwiefern er inhaltlich relevant ist, lässt sich ohne Blick in die konkreten Aufgaben nur teilweise beantworten. Man sollte sicherheitshalber vermutlich ± 1 korrekte Aufgabe als aus eingeschränkter Passung des Rasch-Modells bzw. nicht völlig übereinstimmenden Testheften resultierendes Vertrauensintervall berücksichtigen. Dann verbleibt ein halbwegs sicherer Unterschied von einer Aufgabe, was unabhängig von den eingesetzten Aufgaben als inhaltlich nicht relevant gelten dürfte. Berücksichtigt man zusätzlich, dass Baseline-M8 eine Stichprobenerhebung war und der Test im „Ernstfall“ Vollerhebung 2012 von Lehrenden und Lernenden vermutlich deutlich ernster genommen wurde, so lässt sich der aufgetretene „small effect“ wohl bereits dadurch erklären.

Es ist hier wichtig darauf hinzuweisen, dass die Werte der 500 ± 100-Skala als Punktschätzer grundsätzlich inklusive der Vertrauensintervalle interpretiert werden sollten, insbesondere für kleinere Teilmengen, sei es von Personen (z. B. Schulen oder Schulklassen), sei es von Aufgaben (einzelne Inhalts- oder Handlungsbereiche). In Gesprächen mit dem bifie wurde uns hier ein Richtwert von mindestens $\pm 15 - 25$ Punkten auf der Gesamtskala (90%-Vertrauensintervall) für Schulklassen genannt, unter dem man nicht von einem Unterschied ausgehen kann, auf Einzelschüler(innen)ebene kommen noch einmal etwa 15 Punkte hinzu. Dieser Werte dürften noch eher konservativ sein, für die Subskalen wären sie dann nochmals zu erhöhen.

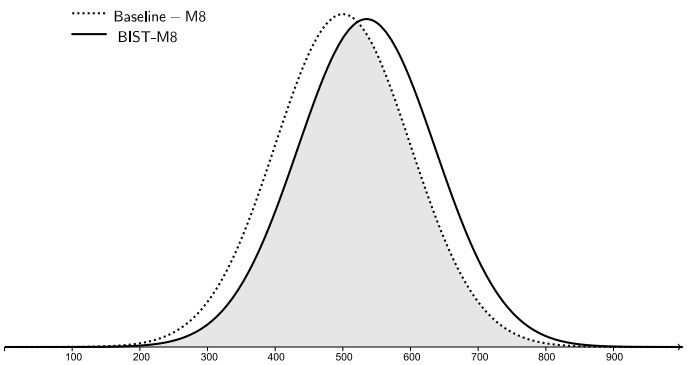

Abb. 4: Fähigkeitswerte Baseline-M8 <> BIST-M8 (eig. Berechnung)

Man sollte außerdem nicht vergessen, dass Mittelwertdifferenzen die Unterschiedlichkeit von Verteilungen stark pointieren. Lenkt man den Blick auf die gesamten Verteilungen (s. Abb. 4), so würden sich diese zu über 85% überlappen[22]. Als Gedankenexperiment: Könnte man in einem Raum 100 typische Baseline-M8 Personen, im Nebenraum 100 typische BIST-M8 Personen versammeln, dann muss man aus beiden Räumen 85 Personen entfernen, wenn der/die Beste im ersten Raum höchstens so gut sein soll, wie der/die Schwächste im zweiten Raum.

4.2. Wie bedeutsam sind Bundesländer-, Schulform- und Geschlechterunterschiede?

Vergleichsgruppe 1		Korrekte Lösungen	Vergleichsgruppe 2		Korrekte Lösungen	Überlappung Gruppe 1 und 2
Wien	(517)	21 (45%)	Oberösterreich	(548)	24 (50%)	88%
APS	(504)	20 (42%)	AHS	(600)	30 (60%)	60%
Buben	(539)	24 (49%)	Mädchen	(532)	23 (47%)	96%

Tab. 6: Gruppenvergleiche (Breit u. Schreiner 2012, eig. Berechnungen[23])

Zur Beantwortung wird erneut auf die Unterschiede in den unskalierten Mittelwerten (Aufgaben korrekt in einem idealtypischen 48 Aufgaben Testheft) und Überlappungen von Verteilungen (s. Tabelle 6) zurückgegangen. Man sollte erneut einen Unsicherheitsbereich von ± 1 Aufgabe berücksichtigen. Der auf der 500 ± 100-Skala gemessene *Unterschied zwischen Buben und Mädchen* verschwindet auf Ebene der gelösten Aufgaben in einem idealtypischen Testheft faktisch vollständig. Manchmal hört man Spekulationen, dieser fehlende Leistungsunterschied sei vermutlich durch eine gezielte Aufgabenauswahl gesteuert worden. Dazu ist zu sagen, dass uniforme Leistungsunterschiede zwischen Teilgruppen (also etwa Buben und Mädchen) vollkommen mit dem Rasch-Modell verträglich wären. Ebenso wie es dem Rasch-Modell zunächst keinen Abbruch tut, dass AHS und APS erkennbar unterscheidlich abgeschnitten haben, wäre dies (falls es dort ähnlich wäre) messtheoretisch auch für Buben und Mädchen zulässig, so lange sich dieser Unterschied auf *alle* Aufgaben bezieht und sich insbesondere die *Schwierigkeitsreihenfolge* der Aufgaben in beiden Gruppen nicht deutlich unterscheidet.

Grundsätzlich unzulässig sind bei einer eindimensionalen Skala aber unterschiedliche relative Leistungsstärken/Leistungsschwächen größerer Teilgruppen. Gäbe es also z. B. 10 Aufgaben, die Buben leicht fallen, Mädchen aber schwer und weitere 10 Aufgaben, die Mädchen schwer fallen, aber Buben leicht, wohingegen sich die Schwierigkeiten bei den übrigen 28 Aufgaben für Mädchen und Buben gleich darstellen, dann sollte dies in der Phase der Pilotierung auffallen und man würde die ersten 20 Aufgaben aufgrund sog. DIF-Effekte („differential item functioning“) ausschließen wollen. Im den uns vorliegenden Testdaten schwanken die Lösungshäufigkeitsunterschiede bei einzelnen Aufgaben zwischen Buben und Mädchen immer noch deutlich stärker als in Summe, wobei es unter den 72 bei 90% der Schüler(innen) eingesetzten Aufgaben vor allem im mittleren Schwierigkeitsbereich jeweils solche gibt (vgl. Abb. 5), bei denen entweder Mädchen (bis zu 8%-Punkte) oder Buben (bis zu 18%-Punkte) besser abschneiden als das jeweils andere Geschlecht[24]. Man dürfte also nicht sehr radikal auf das Ausschließen von DIF-Effekten erzeugende Aufgaben geachtet haben bzw. war die Pilottestung nur mäßig erfolgreich im Aufdecken solcher Aufgaben[25]. Aus fachdidaktischer Sicht ist der Ausschluß von Aufgaben aufgrund solcher Effekte ohnehin nicht unproblematisch, weil sie ebensogut auf mangelnde Eindimensionalität des gemessenen Konstrukts hinweisen können, wie auf Benachteiligung bestimmter Personengruppen (s. Abs. 4.4).

[22] Hier und bei Frage 2 wird jeweils Passung des Rasch-Modells und Zulässigkeit der Normalverteilungsannahme vorausgesetzt.

[23] Prozentwerte der Bundesländer sind aus den 500 ± 100-Werte zurückgerechnete approximative Werte, Prozentwerte für Buben, Mädchen, AHS und APS sind Mittelwerte der rohen Lösungshäufigkeiten der einzelnen Aufgaben, wie sie dem IDM Klagenfurt vorliegen. Korreliert man die Rohwerte für die vier Gruppen mit aus dem Modell zurückberechneten Werten, so ergeben sich Korrelationen von $r > 0{,}98$. Es spricht daher vermutlich wenig dagegen, bei den anderen beiden Gruppen zurückberechnete Werte zu verwenden.

[24] Betrachtet man alle 249 eingesetzten Aufgaben, gibt es noch dramatischere Unterschiede, allerdings werden die Teilgruppen, denen die Aufgaben überhaupt gestellt wurden, teilweise auch sehr klein (unter 500 Personen).

[25] Beides reduziert prinzipiell die Aussagekraft der Skala, insbesondere auf ihr eingeteilter „Kompetenzstufen“.

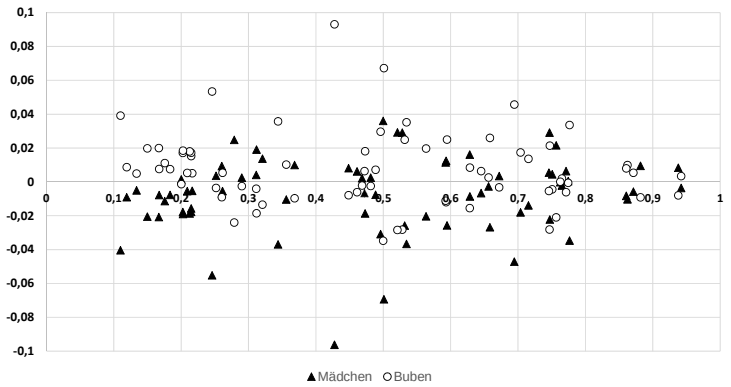

Abb. 5: Abweichungen Geschlechter-Mittelwerte <> Gesamtmittelwert (eig. Berechnung)

Der *Unterschied zwischen Schüler(inne)n der AHS und der APS* ist mit 9 Aufgaben eines idealtypischen Testhefts eine von nur zwei Stellen, an denen sich eine Abschätzung in Anzahlen von Aufgaben auch im Bundesergebnisbericht selbst findet (vgl. Breit u. Schreiner 2012, S. 62, die andere Stelle ist der nicht vorhandene Geschlechterunterschied, a. a. O., S. 30). Tatsächlich dürfte er das vielleicht am wenigsten überraschende Ergebnis der Testung darstellen. Ich würde mich hier vollkommen der bifie-Einschätzung anschließen, dass er ganz wesentlich dem Umstand zu verdanken sein dürfte, dass man in Österreich eben eine Schulsystem hat, in dem eine „leistungsbezogene Trennung vieler Schüler/innen mit 10 Jahren“ stattfindet und „erwartungsgemäß die Gesamtergebnisse der AHS-Schüler/innen in allen Bereichen deutlich besser ausfallen als jene der APS-Schüler/innen“ (a. a. O.). Für die Frage, ob so eine Trennung sinnvoll ist oder nicht, geben die Werte erst einmal nichts her, dafür bräuchte man längsschnittliche Daten, die etwas über die Weiterentwicklung von Fähigkeiten relativ zum Beginn der Unterstufe aussagen. Man sollte auch darauf hinweisen, dass sich die Gesamtverteilungen selbst bei dem deutlichen Mittelwertunterschied immer noch zu 60% überlappen – mehr als die Hälfte der beide Schulformen besuchenden Klientel unterscheidet sich hinsichtlich ihrer durch BIST-M8 gemessenen Mathematikleistung also am Ende der Unterstufe *nicht*.

Bei den *Bundesländerunterschieden* ist der größte Unterschied (Wien <> Oberösterreich) immer noch kleiner, als der Unterschied der Gesamtgruppen zwischen den Messzeitpunkten 2009 und 2012. Sämtliche durch die Medien gehechelten „Ligatabellen“ der Länder spielen sich also im Bereich von Unterschieden von höchstens drei korrekt gelösten Aufgaben mehr oder weniger ab (zur (Nicht-)Relevanz s. oben).

4.3. Ist „Statistische Darstellungen und Kenngrößen“ der „am besten“ absolvierte Inhaltsbereich?

Hier ist einzuräumen, dass das so im Bundesergebnisbericht selbst gar nicht behauptet wird. Diesem Bericht ist lediglich zu entnehmen, dass der *Leistungsfortschritt* seit 2009 im Inhaltsbereich „Statistische Darstellungen und Kenngrößen“ am größten ausgefallen ist. Diese Aussage ist eine Interpretation der Tatsache, dass der Punktwert auf der 500 ± 100-Skala im diesem Inhaltsbereich mit 544 höher liegt, als in allen anderen Subskalen – wobei die Werte aller Subskalen ohnehin nur zwischen 519 und 544 schwanken. Absolute Vergleiche zwischen diesen Werten sind nicht möglich, weil alle Handlungsbereiche 2009 getrennt auf 500 als Mittelwert fixiert wurden. Streng genommen müsste man sagen: Die Effektstärke der Mittelwertunterschiede ist im Inhaltsbereich „Statistische Darstellungen und Kenngrößen“ am größten, also: relativ zur Leistung und zur Streuung in den einzelnen Inhaltsbereichen im Jahr 2009 haben sich die Lösungsquoten hier am deutlichsten positiv entwickelt. Man kann aber noch nicht einmal sagen, dass der Zuwachs an Lösungshäufigkeit in Prozentpunkten hier zwingend am größten gewesen sein muss (weil es 2009 bereits unterschiedliche Streuung je nach Inhaltsbereich gegeben haben kann).

Man kann nun wieder die unskalierten Lösungshäufigkeiten als Kontrollgröße der Effektstärke zu Rate ziehen. Es ergibt sich allerdings ein technisches und ein erkenntnislogisches Problem: Die technische Schwierigkeit besteht darin, dass der dem IDM Klagenfurt zur Verfügung gestellte Datensatz keine Zu-

ordnungen der Aufgaben zu Inhalts- und Handlungsbereichen des Kompetenzmodells enthält. Wir haben daher (uns stehen auch alle Aufgaben zur Verfügung) selbst solche Zuordnungen vorgenommen, diese dürften zumindest für die Inhaltsbereiche mit den Zuordnungen des bifie weitgehend übereinstimmen (vgl. den Beitrag von Schneider in diesem Band). Das schwerer wiegende erkenntnislogische Problem besteht darin, dass wir zwar sehen, dass es in den Inhaltsbereichen zu unterschiedlichen mittleren Lösungshäufigkeiten gekommen ist, dies aber immer zwei Interpretationen zulässt: Die Schüler(innen) beherrschen die in den Inhaltsbereichen angesprochenen Fähigkeiten unterschiedlich gut *oder* in den verschiedenen Inhaltsbereichen sind (bewusst oder unbewusst) per se unterschiedlich schwierige Aufgaben (im Sinne kognitiver Anforderungen) gestellt worden[26]. Ohne Blick auf die Aufgaben selbst und ohne kriteriale Norm im Hintergrund[27] sind aussagekräftige Vergleiche zwischen den Handlungsbereichen im Jahr 2012 sowie der jeweils erzielten Fortschritte nicht möglich. Ohne dem Beitrag von Edith Schneider vorgreifen zu wollen: Tatsächlich ist der Inhaltsbereich „Statistische Darstellungen und Kenngrößen“ der gemäß Lösungshäufigkeit in BIST-M8 (Teilstudie S90) mit $44,9\%$ am schlechtesten absolvierte, gewonnen hätte hier recht deutlich ($52,3\%$) der Bereich „Geometrische Figuren und Körper“, der auf der 500 ± 100-Skala das Schlusslicht bildet. Bei den Handlungsbereichen ergeben sich noch weit dramatischere Unterschiede: Hier liegt „Rechnen, Operieren“ mit $57,3\%$ deutlich vorne, „Argumentieren, Begründen“ mit $26,4\%$ abgeschlagen auf dem letzten Platz. Für die Handlungsbereiche dürften sich allerdings die Zuordnungen der Aufgaben zwischen IDM und bifie vermutlich deutlich voneinander unterscheiden, was einen Vergleich mit den 500 ± 100 Werten nur eingeschränkt möglich macht.

4.4. Wie passen statistische Modellierung und Ergebnisrückmeldung zu Zielsetzungen und theoretischem Kompetenzmodell der Standards?

Ein noch weit größeres Problem bei der Bildung von Subskalen für Inhalts- und Handlungsbereiche besteht allerdings darin, dass hier Unterskalen für ein zunächst eindimensional angenommenes Messmodell gebildet werden. Wenn das Rasch-Modell für die Gesamtheit aller Aufgaben in Baseline-M8 und BIST-M8 gelten soll, diese also alle auf einer eindimensionalen Skala abgebildet werden können, so stellen unterscheidliche Entwicklungen in einzelnen Aufgabengruppen (z. B. Handlungsbereiche) *entweder* irrelevante Messfehler dar, *oder* sie geben Hinweise auf relevante, durch das Modell nicht erklärte Abweichungen, weisen also auf eine nicht modellierte Mehrdimensionalität des zugrunde liegenden Fähigkeitskonstrukts hin. Ich erinnere an die zwei zentralen Grundannahmen des Rasch-Modells:

1. Für die Einschätzung der Fähigkeit einer Person ist nur entscheidend, *wie viele*, aber *nicht welche* Aufgaben sie korrekt gelöst hat.
2. Für die Einschätzung der Schwierigkeit einer Aufgabe ist nur entscheidend, *wie viele*, aber *nicht welche* Personen sie korrekt gelöst haben.

Wenn Aufgaben zu „Statistische Darstellungen und Kenngrößen“ den Personen in BIST-M8 tatsächlich relativ zu den übrigen Aufgaben leichter gefallen sind, als dies bei Baseline-M8 der Fall war, dann wird tendenziell die zweite Grundannahme verletzt: Aufgaben aus dem Bereich „Statistische Darstellungen und Kenngrößen“ werden sich in der Rangliste aller Aufgaben nach unten (zu den leichteren Aufgaben) bewegen, andere dafür nach oben. Für die Einschätzung der Schwierigkeit einer Aufgabe aus diesem Inhaltsbereich wäre es dann nicht mehr egal, ob sie von Personen aus Baseline-M8 oder BIST-M8 gelöst würden, also gerade nicht egal, *welche* Personen sie gelöst hätten. Dadurch entsteht eine zutiefst paradoxe Situation: Fachdidaktisch ist es prinzipiell wünschenswert, dass es in bislang weniger gut laufenden Teilbereichen mit den Lösungshäufigkeiten sowohl absolut als auch relativ zu besser laufenden Teilbereichen nach oben geht. Psychometrisch betrachtet führen solche Veränderungen in der internen Fähigkeitsstuktur aber u. U. dazu, dass eine zuvor empirisch als homogen empfundene Zusammenstellung von Aufgaben als weniger homogen empfunden wird und man bestimmte Aufgaben daher entfernen

[26] Lehrperson kennen den Effekt: Man merkt während des Unterrichts, dass ein bestimmtes Teilthema nicht so gut gelaufen ist und beschränkt sich dort in der Schularbeit dann auf (vermeintlich) „einfachere“ Aufgaben. Das geht in einer zentralen Testung prinzipiell noch einfacher, wo man Aufgaben umfangreich pilotieren kann.

[27] Sind die Anforderungen inhaltlich gerechtfertigt? Müssen die Schüler(innen) das im Sinne der Standards leisten können?

müsste (vgl. Goldstein 1979, S. 217). Damit läuft man Gefahr, *entweder* genau jene relativen Stärken und Schwächen beständig zu reproduzieren, die bei Baseline-M8 bestanden haben, *oder* aber eine immer schlechtere Passung von Modell und Daten in Kauf zu nehmen. Rost (1996, S. 129) stellt in diesem Zusammenhang fest, man könne „die Prüfung des Raschmodells für einen Datensatz auch als Kriterium benutzen, ob es lohnenswert ist, eine *Patternanalyse* [...] vorzunehmen: Eine solche Patternanalyse ist nur sinnvoll, wenn das Modell *nicht* gilt." Hätten also Personen aus BIST-M8 und Baseline-M8 *relevante* unterschiedliche Lösungshäufigkeitsreihenfolgen („Pattern"), gilt das Rasch-Modell nicht. Soll das Rasch-Modell gelten, so können unterschiedliche Lösungshäufhigkeitsreihenfolgen nur *irrelevante* Messfehler sein – damit wird die Berichterstattung über Subskalen-Ergebnisse aber zur Berichterstattung über Messfehler oder zur Berichterstattung auf Basis eines nicht passenden Modells.

Nicht-modellierte Mehrdimensionalität ist auch insofern problematisch, als das theoretische Kompetenzmodell der Standards gerade *keine* eindimensionale Fähigkeitsstruktur beschreibt (vgl. Peschek 2012, S. 27ff, S. 35ff), die für die Herstellung eines homogenen Aufgabensets bei Baseline-M8 schlicht *pragmatisch* unterstellt wurde. Schon für einen einzelnen Messzeitpunkt lassen sich die Grundannahmen des Rasch-Modells *nicht* aus dem Kompetenzmodell der Standards ableiten. Stellt man nun dennoch ein homogenes Aufgabenset durch Pilotierung und Ausschluss schlecht passender Aufgaben her, stößt man zwangsläufig „auf das seit Cronbach & Gleser (1965) diskutierte Bandbreiten-Fidelitäts-Dilemma" (Vohns 2012, S. 348):

> Genau solche Typen von Aufgaben weisen tendenziell eine „schlechte" psychometrische Qualität auf, bei denen Teilgruppen von Schüler(inne)n systematisch schlechtere Ergebnisse erzielen, als es ihre Leistungen bei den übrigen Aufgaben erwarten lassen. Für bislang in der Unterrichtspaxis weniger weit verbreitete, fachdidaktisch aber besonders wünschenswerte Aufgabentypen ist es alles andere als unwahrscheinlich, dass sich genau dieser Effekt einstellt. Die Herausnahme solcher Aufgaben [...] kann dann in der paradoxen Konsequenz münden, die Breite der mathematischen Anforderungen zu Gunsten des bereits besonders stark in der Unterrichtspraxis Etablierten einzuschränken[28] [...].

Selbst wenn man dem bifie zugesteht, sich nach eigenem Bekunden eher in Richtung Bandbreite als in Richtung Fidelität zu orientieren (also: inhaltliche Relevanz höher zu gewichten als die Herstellung von Homogenität der Skala durch Aufgabenausschlüsse), entkommt es dem Dilemma spätestens auf der Ebene der Schul-, Schulklassen- und Einzelschüler(innen)rückmeldungen grundsätzlich nicht: Das Rasch-Modell ist seinem Prinzip nach ein probabilistisches Testmodell. Für große Gruppen von Personen mag gelten, dass die Personen, die in etwa dieselbe Anzahl an Aufgaben korrekt lösen auch in etwa dieselben Aufgaben korrekt lösen, sich daher durch übergreifende Merkmale dieser Aufgaben ihre Fähigkeiten annähernd beschreiben lassen. Eine einzelne Schule, Schulklasse, erst recht ein(e) einzelne(r) Schüler(in) kann aber selbst bei idealer Passung durch die Lösung *vollkommen anderer* Aufgaben zu seinen/ihren Rohpunkten gekommen sein. Dass Schüler(innen) typischerweise bestimmte Aufgaben lösen, ist eine Aussage über eine Teilmenge besonders wahrscheinlich gelöster Aufgaben im Modell. Praktisch hat dann die Menge der tatsächlich gelösten Aufgaben *immer* mehr oder minder große Schnittmengen mit dieser Teilmenge. Kompetenzstufen beruhen auf der Beschreibung einer idealtypischen Aufgabenmenge, je kleiner die betrachtete Teilgruppe von Personen, desto weniger sicher / aussagekräftig ist aber die Beschreibung durch typische Aufgaben. Das gilt erst recht, wenn man glaubt, dass unterschiedliche Schwerpunkte im Unterricht nicht bloß zu zufälligen Abweichungen in den typischerweise gelösten Aufgaben führen, sondern zu systematischen. Je weniger strikt man zudem auf die Durchsetzung des Rasch-Modells setzt, desto wahrscheinlicher werden größere Abweichungen in den typischerweise korrekt gelösten Aufgaben. Je strikter man das Rasch-Modell umsetzt, desto größer ist umgekehrt die Gefahr, die inhaltliche Breite der Standards gar nicht mehr angemessen durch Testaufgaben abzubilden, also etwas zu testen, was mit den *inhaltlichen* Anliegen der Standards nur noch dem Namen nach etwas zu tun hat.

Auf der Ebene einzelner Personen wirkt die Rückmeldung von Punktwerten (mit Unsicherheiten von wenigstens ± 50 Punkten auf den Subskalen) unredlich, welche dieser Person eben nicht konkret sagt, was sie in dem Test geleistet hat, sondern nur, was sie geleistet hätte, wenn sie ein ähnliches Lösungsverhalten

[28] Vgl. dazu auch Goldstein u. Blinkhorn (1982) [Zitat nicht im Original]

wie die meisten anderen Personen an den Tag legen *würde*. Aufgrund der Geheimhaltung der Testaufgaben ist aber keine konkrete Rückmeldung über gelöste und nicht gelöste Aufgaben möglich. Man hängt hier gewissermaßen in einer argumentativen Schleife fest: Weil man das Rasch-Modell einsetzen will, um Testzeitpunkte miteinander zu verknüpfen und verschiedenen Schler(inne)n verschiedene Aufgabensets anzugedeihen, sie aber trotzdem miteinander vergleichen zu können, muss man die Aufgaben geheim halten und weil man die Aufgaben geheim hält, muss man das Rasch-Modell zur Konstruktion von „Kompetenzsstufen“ einsetzen, um überhaupt etwas rückmelden zu können.

5. Rückschau: Standard-Testungen als (abgebrochenes) Turnier?

Ich will zum Abschluss noch einmal auf das Beispiel von Zermelos Schach-Turnier zurückkommen. Zweck des eingesetzten Messmodells für die Spielstärke war es, zu einer eindeutigen Reihenfolge aller Spieler zu gelangen. Mir erscheint der Gedanke, Aufgaben und Schüler(innen), sowie bestimmte Teilgruppen (seien es regionale, seien es Schulformen, Schulen, Klassen innerhalb einer Schule, seien es Geschlechter) fein säuberlich auf einer Rangliste zu platzieren und damit einen Wettkampfs- bzw. Wettbewerbsgedanken im Bildungsystem zu etablieren, dominiert die Entscheidung *für* das Rasch-Modell viel stärker, als fachdidaktisch-inhaltliche Erwägungen für dieses Modell sprächen. Insofern hielte ich es auch *nicht* für einen Schaden, wenn die derzeitige seitens der Ministerin verordnete Zwangspause bei PISA auch zu einem Überdenken der Ausgestaltung der Testung der Standards führen würde.

Für die lokale Unterrichtsentwicklung vor Ort sind Informationen darüber, wo sich meine Klasse im Vergleich zu den Klassen meiner Schule und im Vergleich zu ganz Österreich gemäß einer bestimmten (der Logik des *theoretischen* Kompetenzmodells der Standards gar nicht folgenden) Rangordnung einsortiert, von sehr begrenztem Wert. Lehrer(innen) sollten wissen, an welchen Anforderungen *ihre eigenen Schüler(innen)* ganz konkret scheitern, und nicht mit nichtssagenden Umschreibungen von Kompetenzsstufen abgespeist werden, die ihnen allenfalls sagen, welche Arten von Aufgaben ihre Schüler(innen) unter der Voraussetzung, dass sie sich idealtypisch modellkonform verhalten, dann mit hoher Wahrscheinlichkeit am ehesten gelöst hätten, wenn sie nicht 2012, sondern 2009 zum Test angetreten wären.

Literatur

[Bender 2005] BENDER, Peter: PISA, Kompetenzstufen und Mathematik-Didaktik. In: *Journal für Mathematik-Didaktik* 26 (2005), Nr. 3/4, S. 274–281.

[Breit u. Schreiner 2010] BREIT, Simone; SCHREINER, Claudia (Hrsg.): *Bildungsstandards: Baseline 2009 (8. Schulstufe). Technischer Bericht*. https://www.bifie.at/buch/1116. Stand: 21. März 2014.

[Breit u. Schreiner 2012] BREIT, Simone; SCHREINER, Claudia (Hrsg.): *Bundesergebnisbericht Standardüberprüfung Mathematik 2012, 8. Schulstufe*. https://www.bifie.at/node/1948. Stand: 21. März 2014.

[Bryce 1981] BRYCE, Tom G. K.: Rasch-Fitting. In: *British Educational Research Journal* 7 (1981), Nr. 2, S. 137–153.

[Cohen 1988] COHEN, Jacob: *Statistical power analysis for the behavioral sciences*. 2. Aufl., Hillsdale, NJ: Erlbaum, 1988.

[Cronbach u. Gleser 1965] CRONBACH, L. J.; GLESER, G. C.: *Psychological tests and personnel decisions*. Urbana: University of Illinois Press, 1965.

[Freunberger 2013] FREUNBERGER, Roman: *Standard-Setting Mathematik 8. Schulstufe. Technischer Bericht*. https://www.bifie.at/node/2192. Stand: 21. März 2014.

[Goldstein 1979] GOLDSTEIN, Harvey: Consequences of Using the Rasch Model for Educational Assessment. In: *British Educational Research Journal* 5 (1979), Nr. 2, S. 211–220.

[Goldstein 1980] GOLDSTEIN, Harvey: A Rejoinder to Preece. In: *British Educational Research Journal* 6 (1980), Nr. 2, S. 211–212.

[Goldstein u. Blinkhorn 1982] GOLDSTEIN, Harvey; BLINKHORN, Steve: The Rasch Model Still Does Not Fit. In: *British Educational Research Journal* 8 (1982), Nr. 2, S. 167–170.

[Hattie 2009] HATTIE, John: *Visible Learning*. London: Routledge, 2009.

[IDM 2007] IDM, Institut für Didaktik der Mathematik: *Standards für die mathematischen Fähigkeiten österreichischer Schülerinnen und Schüler am Ende der 8. Schulstufe. Version 4/07*. http://www.uni-klu.ac.at/idm/downloads/Standardkonzept_Version_4-07.pdf. Stand: 22. März 2014.

[Lind u. a. 2005] LIND, Detlef; KNOCHE, Norbert; BLUM, Werner; NEUBRAND, Michael: Kompetenzstufen in PISA. In: *Journal für Mathematik-Didaktik* 26 (2005), Nr. 1, S. 80–87.

[Meyerhöfer 2004] MEYERHÖFER, Wolfram: Zum Kompetenzstufenmodell von PISA. In: *Journal für Mathematik-Didaktik* 25 (2004), Nr. 3/4, S. 294–305.

[Nimmervoll 2014a] NIMMERVOLL, Lisa: *„Es war nicht zu erwarten, dass die Neuen Mittelschulen alle anderen überflügeln" (derstandard.at, 10.02.2014)*. http://goo.gl/EMz6uo. Stand: 22. März 2014.

[Nimmervoll 2014b] NIMMERVOLL, Lisa: *„Bifie-Aufgaben werden wie der Kronschatz gehütet" (derstandard.at, 16.02.2014)*. http://goo.gl/Ral8Ef. Stand: 22. März 2014.

[Peschek 2012] PESCHEK, Werner: Die österreichischen Standards M8. In: KRÖPFL, Bernhard; SCHNEIDER, Edith (Hrsg.): *Standards Mathematik unter der Lupe*. München; Wien: Profil, 2012, S. 21–40.

[Preece 1980] PREECE, Peter F. W.: On Rashly Rejecting Rasch: A Response to Goldstein (With a Rejoinder from Goldstein). In: *British Educational Research Journal* 6 (1980), Nr. 2, S. 209–211.

[Ratzka 2004] RATZKA, Nadja: Mathematische Leistung im Spiegel unterschiedlicher Tests. In: ESSLINGER-HINZ, I.; HAHN, H. (Hrsg.): *Kompetenzen entwickeln – Unterrichtsqualität in der Grundschule steigern*. Hohengehren: Schneider Verlag, 2004, S. 175–179.

[Rost 1996] ROST, Jürgen: *Lehrbuch Testtheorie Testkonstruktion*. Bern: Hans Huber, 1996.

[Rost 2004] ROST, Jürgen: Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. In: *Zeitschrift für Pädagogik* 50 (2004), Nr. 5, S. 662–678.

[Strobl 2010] STROBL, Carolin: *Das Rasch-Modell. Eine verständliche Einführung für Studium und Praxis*. München u. Mering: Rainer Hampp Verlag, 2010.

[Vohns 2012] VOHNS, Andreas: Zur Rekonstruierbarkeit impliziter Standardsetzungen zentraler Prüfungen mit Hilfe des Rasch-Modells. In: *Journal für Mathematik-Didaktik* 33 (2012), Nr. 2, S. 339–349.

[Wuttke 2007] WUTTKE, Joachim: Die Insignifikanz signifikanter Unterschiede: der Genauigkeitsanspruch von PISA ist illusorisch. In: JAHNKE, Thomas; MEYERHÖFER, Wolfram (Hrsg.): *PISA & Co : Kritik eines Programms*. Hildesheim; Berlin: franzbecker, 2007, S. 99–246.

[Zermelo 1929] ZERMELO, Ernst: Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. In: *Mathematische Zeitschrift* 29 (1929), Nr. 1, S. 436–460.

Anschrift des Verfassers

Assoz. Prof. Dr. Andreas Vohns
Alpen-Adria-Universität Klagenfurt
Institut für Didaktik der Mathematik
Sterneckstraße 15
9010 Klagenfurt
Email: andreas.vohns@aau.at