Vergleich von Mathematikkompetenzen zwischen den Klassenstufen 5 und 7 mit NEPS-Daten

G. Rohwer (September 2016)

Zusammenfassung Der Beitrag diskutiert, wie man mit Daten des Nationalen Bildungspanels (NEPS) Veränderungen von Mathematikkompetenzen zwischen den Klassenstufen 5 und 7 erfassen kann. Verwendet werden Ergebnisse aus zwei Tests mit 23 bzw. 22 binären Items, von denen sechs in beiden Tests gleich sind. Der Beitrag argumentiert, dass für einen Vergleich ein beide Tests umfassendes Raschmodell sowohl aus statistischen als auch aus theoretischen Gründen nicht geeignet ist. Das statistische Argument bezieht sich darauf, dass die Annahme zeitinvarianter Itemparameter abgelehnt werden muss. Das theoretische Argument bezieht sich darauf, dass mit einem solchen Modell nicht adäquat berücksichtigt werden kann, dass Schüler lernen, neue Arten von Mathematikaufgaben zu lösen. Um dies zu berücksichtigen, wird vorgeschlagen, sich gedanklich auf einen Test zu beziehen, der alle in den beiden Tests verwendeten Items umfasst. Dafür liefern jedoch die verfügbaren Testergebnisse nur unzureichende Informationen. Es wird gezeigt, dass daraus erhebliche Fehler bei der Einschätzung von Kompetenzveränderungen resultieren können. Schließlich wird ausgeführt, dass man verlässliche Aussagen nur aus den in beiden Tests gemeinsam verwendeten Items gewinnen kann.

Schlüsselwörter Mathematische Kompetenz · Kompetenzveränderungen · Raschmodell · NEPS-Daten

Using NEPS Data for Comparing Math Competencies Between Grade 5 and Grade 7

Abstract The article discusses how to use data of the National Educational Panel Study (NEPS) for investigating changes of math competencies between grade 5 and grade 7. The data result from two tests consisting, respectively, of 23 and 22 binary items of which six are identical in both tests. It is argued that a joint Rasch model does not provide a suitable framework for mainly two reasons. A statistical argument shows that the assumption of time-invariant item parameters, which is required by the joint Rasch model, must be rejected. A further theoretical argument shows that this model cannot adequately take into account processes in which pupils learn to solve new kinds of mathematical tasks. It is argued that this requires the reference to a comprehensive test which includes all items of the two tests. The actually available tests results are, however, insufficient for an assessment of changes of competencies defined by such a comprehensive test. It is shown that this can lead to heavily biased conclusions. The article therefore proposes that reliable assessments of competence changes can only be based on the items used in both tests.

Keywords Math competence · Changes of competencies · Rasch model · NEPS data

1 Einleitung

Dieser Beitrag diskutiert, wie man mit Daten des Nationalen Bildungspanels (NEPS) Veränderungen von Mathematikkompetenzen zwischen den Klassenstufen 5 und 7 erfassen kann. Dafür beziehe ich mich auf Ergebnisse von zwei Tests, die in der Startkohorte Klasse 5 in diesen Klassenstufen durchgeführt wurden (Duchhardt und Gerdes 2012; Neumann et al. 2013).[1] 5194 Schüler haben am ersten Test in der Klasse 5 teilgenommen und mindestens eine gültige Antwort gegeben. 3833 von ihnen haben auch am zweiten Test in der Klasse 7 teilgenommen und mindestens eine gültige Antwort gegeben; dies ist die Anzahl der Schüler, die in diesem Beitrag betrachtet werden.

Der Test in der Klasse 5 besteht aus 23 binären und einem komplexen Item, der Test in der Klasse 7 besteht aus 22 binären und einem komplexen Item. Sechs binäre Items sind in beiden Tests identisch. Um die Diskussion und Notation zu vereinfachen, verwende ich nur die binären Items und unterscheide nur zwischen richtigen und nicht richtigen Antworten.

[1] Diese Arbeit nutzt Daten des Nationalen Bildungspanels (NEPS): Startkohorte Klasse 5, doi:10.5157/NEPS:SC3:4.0.0. Die Daten des NEPS wurden von 2008 bis 2013 als Teil des Rahmenprogramms zur Förderung der empirischen Bildungsforschung erhoben, welches vom Bundesministerium für Bildung und Forschung (BMBF) finanziert wurde. Seit 2014 wird NEPS vom Leibniz-Institut für Bildungsverläufe e.V. (LIfBi) an der Otto-Friedrich-Universität Bamberg in Kooperation mit einem deutschlandweiten Netzwerk weitergeführt. Zur Einführung vgl. man Blossfeld, Roßbach und von Maurice, 2011.

In Abschnitt 2 wird besprochen, ob sich ein gemeinsames Raschmodell für einen Vergleich der Testergebnisse eignet. Ich zeige, dass die dafür erforderliche Annahme zeitinvarianter Itemparameter in diesem Anwendungsfall nicht zutrifft. Dann argumentiere ich, dass diese Annahme auch generell unplausibel ist, weil sie unrealistische Implikationen für die Lernprozesse hat, durch die sich Kompetenzen entwickeln.

In Abschnitt 3 argumentiere ich, dass Testergebnisse auch ohne die Voraussetzung zeitinvarianter Itemparameter verglichen werden können, wenn man sich auf Tests bezieht, die die gleichen Items verwenden. Daraus folgt, dass man sich in unserem Anwendungsfall gedanklich auf einen Test beziehen sollte, der die Gesamtheit der in den beiden Tests verwendeten Items umfasst.

In Abschnitt 4 wird überlegt, ob man sich mit den verfügbaren Informationen auf einen solchen zusammenfassenden Test beziehen kann. Ich kritisiere zunächst die Vorstellung, dass das durch die Annahme eines 'eindimensionalen Konstrukts' gelingen könnte. Dann argumentiere ich, dass man in unserem Anwendungsfall davon ausgehen sollte, dass der Test für die Klasse 7 Items enthält, die Schüler in der Klasse 5 noch nicht lösen konnten. Schließlich zeige ich, dass infolgedessen Testvergleiche durch Gleichsetzung von Itemparametern zu erheblichen Fehlern in der Einschätzung von Kompetenzveränderungen führen können.

Allerdings kann man aus den Ergebnissen der beiden Tests keine sicheren Einsichten in Veränderungen von Kompetenzen gewinnen, wenn diese durch eine Bezugnahme auf alle verwendeten Items definiert werden. Deshalb beschränke ich mich im Abschnitt 5 auf die gemeinsamen Items. Ich argumentiere, dass man zur Quantifizierung wahlweise Summenscores oder durch ein Raschmodell definierte Theta-Werte verwenden kann. Bei einem Vergleich von Kompetenzen zwischen zwei Zeitpunkten entfällt jedoch die Möglichkeit, Summenscores durch ein Raschmodell zu begründen. Schließlich verwende ich Summenscores, um zu zeigen, wie sich durch gemeinsame Items definierte Mathematikkompetenzen zwischen den Klassen 5 und 7 (und ergänzend auch kohortenübergreifend zwischen den Klassen 7 und 9) verändert haben. Der Beitrag endet mit einer kurzen Zusammenfassung.

Notationen: T_1 mit der Itemmenge $J_1 = J^a \cup J^c$ bezeichnet den Test in der Klasse 5, T_2 mit der Itemmenge $J_2 = J^b \cup J^c$ bezeichnet den Test in der Klasse 7. J^c ist die Menge der gemeinsamen Items. Die Variablen (Vektoren) für die (tatsächlichen oder möglichen) Testergebnisse des Tests T_k sind X_k^a, X_k^b und X_k^c $(k = 1, 2)$. Werte der Variablen werden durch entsprechende Kleinbuchstaben bezeichnet: $x_{i,k}^a$, $x_{i,k}^b$ und $x_{i,k}^c$ für $i = 1, \ldots, n = 3833$; itemspezifische Komponenten werden zusätzlich durch einen Index j kenntlich gemacht. Für X_2^a und X_1^b gibt es keine beobachteten Werte.

2 Kann man ein Raschmodell verwenden?

2.1 Das Raschmodell für zwei Zeitpunkte

Ich betrachte im Folgenden zunächst nur die sechs gemeinsamen Items. Für die Klassenstufe k kann ein Raschmodell für eine Person i folgendermaßen geschrieben werden:

$$\mathrm{P}(X_k^c = x_{i,k}^c \,|\, \theta_{i,k}^c, \delta_k^c) = \prod_{j \in J^c} \frac{\exp(\theta_{i,k}^c - \delta_{j,k}^c)^{x_{ij,k}^c}}{1 + \exp(\theta_{i,k}^c - \delta_{j,k}^c)} \tag{1}$$

Auf der linken Seite steht die Wahrscheinlichkeit, dass die Person den Antwortvektor $x_{i,k}^c$ erzeugt, wenn die hinter dem Bedingungsstrich angeführten Parameter gegeben sind. $\theta_{i,k}^c$ repräsentiert die Mathematikkompetenz der Person zum Zeitpunkt des Tests T_k; die Itemparameter werden durch $\delta_{j,k}^c$ bezeichnet und im Vektor δ_k^c zusammengefasst. Zur Identifikation der Modellparameter verwende ich $\sum_{j \in J^c} \delta_{j,k}^c = 0$ (für $k = 1, 2$).

Das Modell impliziert folgenden Anspruch: Wenn die Itemparameter δ_k^c gegeben sind, hängen die Testergebnisse einer Person nur von dem Theta-Wert ab, den sie zum Zeitpunkt des Tests hatte. Für ein Raschmodell, das sich auf die beiden Tests T_1 und T_2 bezieht, folgt daraus:

$$\begin{aligned} &\mathrm{P}(X_1^c = x_{i,1}^c, X_2^c = x_{i,2}^c \,|\, \theta_{i,1}^c, \delta_1^c, \theta_{i,2}^c, \delta_2^c) = \\ &\quad \mathrm{P}(X_1^c = x_{i,1}^c \,|\, \theta_{i,1}^c, \delta_1^c)\, \mathrm{P}(X_2^c = x_{i,2}^c \,|\, \theta_{i,2}^c, \delta_2^c) = \\ &\quad\quad \prod_{j \in J^c} \frac{\exp(\theta_{i,1}^c - \delta_{j,1}^c)^{x_{ij,1}^c}}{1 + \exp(\theta_{i,1}^c - \delta_{j,1}^c)} \prod_{j \in J^c} \frac{\exp(\theta_{i,2}^c - \delta_{j,2}^c)^{x_{ij,2}^c}}{1 + \exp(\theta_{i,2}^c - \delta_{j,2}^c)} \end{aligned} \tag{2}$$

Die Raschmodelle für T_1 und T_2 können also separat geschätzt werden. Damit die geschätzten Theta-Werte zwischen den beiden Tests verglichen werden kön-

nen, wird jedoch gefordert, dass sich die Itemparameter nicht verändern (vgl. z.B. von Davier und von Davier 2007; Fischer et al. 2016), was als eine Bedingung für die gemeinsame Schätzung der beiden Modellteile formuliert werden kann.

2.2 Zeitlich invariante Itemparameter?

Ist diese Forderung mit den Daten vereinbar? Das kann mit einem Likelihood-Ratio-Test untersucht werden. Wenn man das Modell (2) zunächst ohne einschränkende Bedingungen schätzt,[2] findet man die Log-Likelihood: $-6204.3 - 5390.1 = -11594.4$. Schätzt man dann das Modell unter der Annahme gleicher Itemparameter ($\delta^c_{j,1} = \delta^c_{j,2}$ für $j \in J^c$), findet man die Log-Likelihood -11637.8. Die Teststatistik hat also den Wert 86.8 und zeigt bei 5 Freiheitsgraden, dass die Annahme zeitlich konstanter Itemparameter verworfen werden sollte.

Dies Ergebnis illustriert auch einen weiteren wichtigen Punkt: Die Itemparameter eines Raschmodells charakterisieren nicht Eigenschaften von Items (denn diese sind ja in den beiden Tests identisch), sondern sie reflektieren die Verteilung von Kompetenzen in einer Population. Das ist unproblematisch, wenn man das Raschmodell nur für die Erfassung von Kompetenzen zu einem Zeitpunkt verwendet. Sobald man sich aber auf zwei Zeitpunkte bezieht, muss man annehmen, dass sich die Verteilungen der Kompetenzen – und infolgedessen auch die Itemparameter – zwischen den Zeitpunkten verändern.

Man kann auch leicht weitere Beispiele für sich verändernde Itemparameter finden. Zum Beispiel zwei Tests von Mathematikkompetenzen, die im zweiten Teil der BiKS-Studie in den Wellen 4 und 5 durchgeführt wurden.[3] Ich betrachte 1344 Schüler, die an beiden Tests teilgenommen haben. Es gibt 8 gemeinsame Items. Ein gemeinsames Raschmodell ohne weitere Bedingungen

[2] Ich verwende hier und im Folgenden die CML-Methode ('conditional maximum likelihood'), weil man dafür keine Annahmen über die Verteilung der Theta-Werte benötigt. Dagegen wäre die Unterstellung einer zweidimensionalen Normalverteilung für Theta-Werte, wie beispielsweise von Andersen (1985) vorgeschlagen wurde, sehr restriktiv und kaum mit der Unabhängigkeitsforderung (2) vereinbar. Im Folgenden weren also konditionale Log-Likelihoods berichtet, die für die Berechnung der Teststatistik jedoch ausreichend sind.

[3] Artelt, C., Blossfeld, H.-P., Faust, G., Roßbach, H.-G., Weinert, S. (2013): *Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter (BiKS-8-14)*. Version: 1. IQB – Institut zur Qualitätsentwicklung im Bildungswesen. Datensatz: `http://doi.org/10.5159/IQB_BIKS_8_14_v1`.

ergibt die Log-Likelihood -6412.61, und wenn man zeitinvariante Itemparameter fordert, erhält man -6567.93. Die Teststatistik hat also den Wert 310.64 und zeigt offenbar, dass die Annahme zeitinvarianter Itemparameter verworfen werden sollte.

Als ein weiteres Beispiel betrachte ich zwei Tests des passiven Wortschatzes, die in der ersten und dritten Welle der NEPS-Startkohorte 2 (Kindergarten) durchgeführt wurden.[4] Am ersten Test im Jahr 2011 haben 2948 Kinder teilgenommen, die zu diesem Zeitpunkt einen Kindergarten besuchten. Zwei Jahre später nahmen 551 dieser Kinder an einem zweiten Test teil. Ich verwende Testergebnisse für 38 gemeinsame Items von 518 Kindern, die an beiden Tests teilgenommen und mindestens eine gültige Antwort gegeben haben. Die Teststatistik hat den Wert 357, und mit 37 Freiheitsgraden muss die Annahme gleichbleibender Itemparameter auch in diesem Beispiel verworfen werden.

2.3 Lernprozesse und Itemparameter

Es sprechen auch theoretische Überlegungen gegen die Annahme zeitinvarianter Itemparameter, denn sie hat sehr restriktive und sehr unplausible Implikationen für die Lernprozesse, durch die sich Kompetenzen bilden und verändern.

Das kann man sich folgendermaßen verdeutlichen. Es sei $\theta(t)$ der durch ein Raschmodell postulierte Theta-Wert, vorgestellt als eine Funktion der Zeit t. Die zeitabhängige Wahrscheinlichkeit für eine korrekte Antwort für ein Item j mit dem Parameter δ_j ist

$$\pi_j(t) = \frac{\exp(\theta(t) - \delta_j)}{1 + \exp(\theta(t) - \delta_j)} \tag{3}$$

Für zwei Items, $j = 1, 2$, folgt daraus die Beziehung

$$\pi_2(t) = \frac{\pi_1(t)}{c + (1 - c)\,\pi_1(t)} \quad \text{wobei} \quad c = \exp(\delta_2 - \delta_1)$$

Eine Längsschnittvariante des Raschmodells verlangt also strikte deterministische Beziehungen zwischen den Prozessen, durch die itemspezifische Kompetenzen erlernt werden und sich verändern.

Zur Illustration betrachte ich zwei Items mit $\delta_1 = 0$ und $\delta_2 = 1$. Die Zeitachse läuft von 0 bis 10. Die durchgezogene Linie in Abb. 1 zeigt eine willkürlich angenommene Lernkurve für Item 1: $0.8\,\exp(t-3)/(1+\exp(t-3))$.

[4] Ich verwende die Version `doi:10.5157/NEPS:SC2:3.0.0`; siehe auch Fußnote 1.

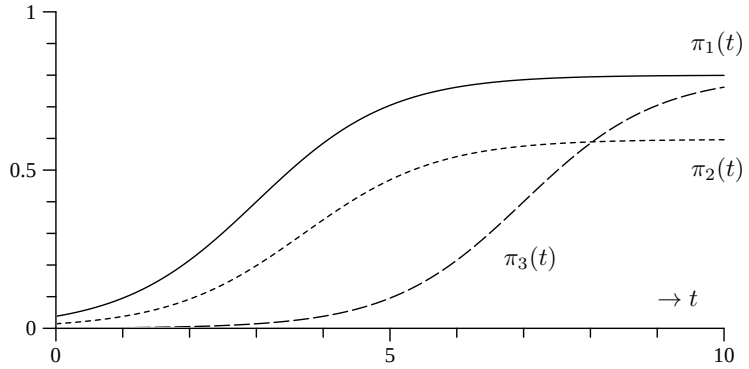

Abb. 1 Drei Lernkurven entsprechend der Definition (3).

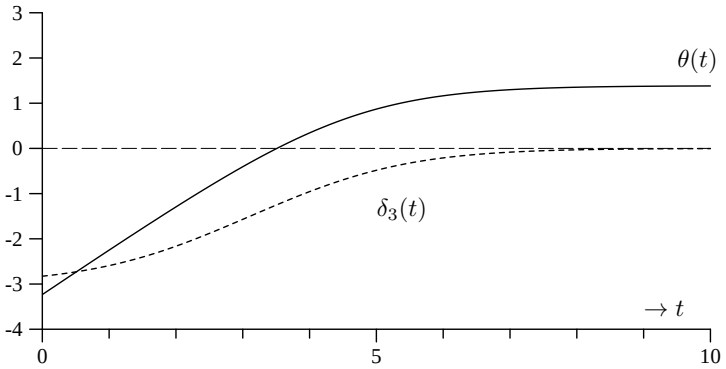

Abb. 2 $\theta(t)$ entspricht der Lernkurve $\pi_1(t)$, $\delta_3(t)$ ist der Parameter von Item 3 entsprechend der Gleichung (4).

Wenn ein Raschmodell gilt, kann man die Lernkurve $\pi_2(t)$ für Item 2 deterministisch ableiten. Andererseits sind jedoch fast alle anderen Lernkurven nicht mit einem Raschmodell vereinbar, das für Item 1 gilt.

Als Beispiel betrachte ich die Lernkurve $\pi_3(t) = \pi_1(t-4)$, die mit $\pi_1(t)$ bis auf eine zeitliche Verschiebung identisch ist. $\theta(t)$ sei die durch die Lernkurve $\pi_1(t)$ implizierte Kompetenzentwicklung. Verwendet man sie als Referenz, findet man für den Parameter von Item 3:

$$\delta_3(t) = \delta_1 + \theta(t) - \theta(t+4) \tag{4}$$

Wie in Abb. 2 illustriert wird, muss sich dieser Parameter im Zeitablauf verändern. Das Beispiel zeigt, dass bereits kleine zeitliche Verschiebungen im Timing von itemspezifischen Lernprozessen nicht mit der Annahme zeitinvarianter Itemparameter vereinbar sind.

3 Definitionen vergleichbarer Kompetenzen

Einige Autoren haben behauptet, dass zeitinvariante Itemparameter eine notwendige Voraussetzung für die Annahme sind, dass zwei Tests 'das gleiche Konstrukt' erfassen (z.B. Stocking und Lord 1983; Rupp und Zumbo 2006; Millsap 2010). In unserem Anwendungsfall hätte diese Auffassung die Konsequenz, dass die sechs identischen Items der beiden Tests unterschiedliche Arten mathematischer Kompetenz erfassen.

Es gibt zwei Möglichkeiten, um diese Konsequenz zu vermeiden. Man kann versuchen zu argumentieren, dass es genügt, wenn die Annahme zeitinvarianter Itemparameter 'näherungsweise' gilt, so dass man die Ergebnisse statistischer Tests ignorieren kann. Dieser Weg wurde von Fischer et al. (2016) vorgeschlagen. Eine Alternative, die ich im Weiteren verfolgen möchte, besteht darin, einen Ansatz zur Definition und Quantifizierung von Kompetenzen zu verwenden, der ohne die Annahme zeitinvarianter Itemparameter auskommt. Als Ausgangspunkt verwende ich folgendes Prinzip:

> Eine hinreichende Bedingung dafür, dass zwei Tests eine vergleichbare Kompetenz erfassen, besteht darin, dass sie die gleichen Items verwenden.

Dies ist keine notwendige Bedingung. Man kann sich vorstellen, dass eine bestimmte Art von Kompetenz durch einen sehr umfangreichen Itempool definiert wird. Dann könnte man auch mit zufällig ausgewählen Items Kompetenzen, die durch den Itempool definiert sind, erfassen und vergleichen. Dieser Ansatz wird allerdings im NEPS-Projekt nicht verwendet, sondern es werden für jede zu erfassende Kompetenz gezielt dafür geeignete Items konstruiert.

Das oben genannte Prinzip entspricht der Idee, dass eine 'Art der Kompetenz' eine bestimmte Bedeutung aus den Items gewinnt, die man für ihre Erfassung einsetzt. Als Kontrast kann man folgendes Prinzip betrachten: Zwei Tests erfassen vergleichbare Kompetenzen, wenn man die Parameter der jeweils verwendeten Items gleichsetzen kann. Dies ist jedoch ein rein formales Prinzip, bei dem von der inhaltlichen Bedeutung der Items abstrahiert wird.

Wenn man diesem Prinzip folgt, könnte man die beiden Mathematiktests dadurch vergleichbar machen, dass man anstelle der sechs gemeinsamen Items, deren Parameter sich verändern, sechs andere Items auswählt, deren Parameter sich nicht verändern; zum Beispiel:

gemeinsame Items			eine andere Itemauswahl			
MAG5Q301	0.79	0.80	MAG5D041	0.92	MAG9Q071	0.92
MAG5D051	−2.45	−2.19	MAG5Q221	−0.61	MAG7R081	−0.60
MAG5D052	−0.36	−0.85	MAG5D052	1.42	MAG9V011	1.41
MAG5R251	0.53	0.68	MAG5Q121	1.08	MAG7D061	1.09
MAG5V321	1.41	1.54	MAG5Q131	1.95	MAG7D011	1.94
MAG5R191	0.07	0.02	MAG5V321	0.59	MAG9Q181	0.59

Auf der linken Seite befinden sich die Itemparameter der sechs gemeinsamen Items, die mit dem Modell (2) geschätzt wurden und sich offenbar unterscheiden. Die Items auf der rechten Seite wurden aus einem gemeinsamen Raschmodell für alle Items der beiden Tests ausgewählt (jeweils sechs Items aus dem ersten und aus dem zweiten Test). Die inhaltliche Bedeutung der Items ist in diesem Fall unterschiedlich, aber die Itemparameter sind nahezu identisch.

Noch ein weiteres Beispiel kann zeigen, dass eine rein formale Betrachtung nicht sinnvoll ist. In diesem Beispiel verwende ich ein gemeinsames Raschmodell für 23 Items des Mathematiktests in der Klassenstufe 5 und 22 Items eines Lesetests, der in der gleichen Klassenstufe mit den gleichen Schülern durchgeführt wurde. Wiederum wähle ich jeweils sechs Items mit gleichen Itemparametern aus:

Mathematiktest		Lesetest	
MAG5Q131	1.95	REG70520	1.95
MAG5Q221	−0.61	REG70360	−0.62
MAG5D052	1.42	REG70120	1.42
MAG5D02S	0.26	REG70620	0.27
MAG5D023	−0.85	REG70440	−0.85
MAG5R191	0.78	REG70650	0.78

Folgte man dem rein formalen Prinzip, könnte man behaupten, dass die beiden Itemmengen ‘das gleiche Konstrukt’ erfassen. Aus meiner Sicht liefert jedoch das Beispiel ein Argument dafür, dass man mit dem rein formalen Prinzip nicht zwischen unterschiedlichen Arten von Kompetenzen unterscheiden kann

und dass es deshalb keine Grundlage liefert, um ‘vergleichbare Kompetenzen’ zu definieren.

4 Tests mit unterschiedlichen Itemmengen

Ich nehme im Weiteren an, dass die Verwendung eines Konstrukts zur Erfassung und für einen Vergleich von Kompetenzen es erfordert, dass seine inhaltliche Bedeutung expliziert werden kann. Um eine solche Erläuterung zu vermitteln, muss man sich auf die Items beziehen, die die zu erfassende Kompetenz exemplifizieren. Daraus folgt: Konstrukte, die durch Tests mit unterschiedlichen Items definiert werden, sind unterschiedliche Konstrukte; und dies gilt unabhängig davon, ob die Itemparameter eines statistischen Modells gleichgesetzt werden können.

Das Argument schließt es nicht aus, dass man für zwei Tests T_1 und T_2 ein gemeinsames Konstrukt definieren kann. Aber um diesem Konstrukt eine Bedeutung zu geben, muss man sich auf einen Test T beziehen, der T_1 und T_2 umfasst. Dies kann dadurch erreicht werden, dass man T durch die Gesamtheit der in T_1 und T_2 verwendeten Items definiert. In unserem Anwendungsfall besteht dieser Test also aus der Itemmenge $J = J_1 \cup J_2$.

Die Mathematikkompetenzen in den Klassen 5 und 7 sollten also durch eine Bezugnahme auf diesen gemeinsamen Test erfasst und verglichen werden. Und die Frage ist, ob bzw. wie das mit den Informationen, die aus den beiden einzelnen Tests verfügbar sind, erreicht werden kann.

4.1 Unterstellung eines eindimensionalen Konstrukts?

Zurvor möchte ich kurz eine statistisch motivierte Idee besprechen, die sich so beschreiben lässt: Die Ergebnisse der Tests T_1 und T_2 können verglichen werden, wenn sie sich durch ein eindimensionales Konstrukt statistisch erklären lassen. Diese Idee ergänzt den schon besprochenen Ansatz, zwei Tests durch eine Gleichsetzung der Parameter ihrer Items vergleichbar zu machen. Es handelt sich aber auch um eine Erweiterung, weil es nicht mehr erforderlich erscheint, dass die beiden Tests aus den gleichen Items bestehen. Stattdessen genügt es, dass sich die Itemmengen überschneiden und dass man die Parameter der gemeinsamen Items gleichsetzen kann.

Die Idee ist plausibel, wenn man sich im Querschnitt auf zwei Stichproben

aus derselben Population beziehen kann. In diesem Fall kann man annehmen, dass es in der Population eine bestimmte Verteilung der Kompetenzen gibt und infolgedessen bestimmte Parameter für alle Items in $J_1 \cup J_2$. Somit kann man auch annehmen, dass beide Tests Informationen über die gleichen Itemparameter liefern, und dies liefert schließlich eine Rechtfertigung dafür, die Tests durch die Parameter ihrer gemeinsamen Items vergleichbar zu machen.

All diese Voraussetzungen sind jedoch in längsschnittlichen Anwendungen nicht mehr gegeben. Wie ich bereits ausgeführt habe, muss man dann davon ausgehen, dass sich Itemparameter verändern können, und da sich die Itemmengen nur teilweise überschneiden, hat man in unserem Anwendungsfall folgende Situation:

	Klasse 5	Klasse 7
Items nicht in T_2	δ_1^a	δ_2^a ?
Items in T_1 und T_2	δ_1^c	δ_2^c
Items nnicht in T_1	δ_1^b ?	δ_2^b

Die Gleichsetzung verläuft über die Parameter der gemeinsamen Items: $\delta_1^c = \delta_2^c$. Dies kann, wie bereits besprochen wurde, geprüft werden. In unserem Anwendungsfall sollte die Gleichsetzung zwar aus statistischen Gründen abgelehnt werden;[5] aber selbst wenn man sie 'näherungsweise' akzeptiert, benötigt man noch zwei weitere Gleichsetzungen:

$$\delta_1^a = \delta_2^a \quad \text{und} \quad \delta_1^b = \delta_2^b \tag{5}$$

Ob diese Gleichsetzungen zulässig sind, kann jedoch mit den verfügbaren Daten nicht geprüft werden. Und noch wichtiger: Ein gemeinsames Raschmodell für die beobachteten Testergebnisse ist mit beliebigen Annahmen über die

[5] Das folgt schon aus dem in Abschnitt 2.2 angeführten Testergebnis. Wenn man den gleichen Test mit einem Raschmodell durchführt, das alle Items beider Tests enthält, findet man die Teststatistik 79.2, wiederum mit 5 Freiheitsgraden.

im obigen Schema mit einem Fragezeichen versehenen Itemparameter vereinbar. Somit kann mit dieser Methode nicht einmal in einem rein formalen Sinn die Unterstellung eines eindimensionalen Konstrukts für die beiden Tests begründet werden.

Wäre es sinnvoll, gleichwohl anzunehmen, dass die beiden Tests Informationen über ein eindimensionales Konstrukt liefern? Die folgende Grafik zeigt die Verteilung der mit einem gemeinsamen Raschmodell geschätzten Itemparameter (die für die Gleichsetzung verwendeten gemeinsamen Items sind durch gestrichelte Linien gekennzeichnet):

Aus der Annahme, dass es ein eindimensionales Konstrukt gibt, würde folgen, dass Schüler bereits in der Klasse 5 Items aus dem Test für die Klasse 7 lösen können; und zwar entsprechend dem Theta-Wert, den sie in der Klasse 5 erzielt haben.

Um das Problem zu illustrieren, betrachte ich die jeweils schwierigsten Items: `MAG5Q121` im Test T_1 ($\delta_{15,1} = 1.03$) und `MAG9V091` im Test T_2 ($\delta_{22,2} = 1.61$). Angenommen, ein Schüler hat in der Klasse 5 gelernt, das erste dieser beiden Items mit der Wahrscheinlichkeit 0.9 zu lösen. Auf der Grundlage des Modells folgt daraus der Theta-Wert $\theta_{i,1} = 3.23$. Die Unterstellung eines eindimensionalen Konstrukts würde implizieren, dass dieser Schüler bereits in der Klasse 5 das schwierigste Item des Tests für die Klasse 7 mit der Wahrscheinlichkeit 0.83 lösen kann. Diese Konsequenz erscheint sehr unplausibel, wenn man daran denkt, wie sich Mathematikkompetenzen entwickeln, insbesondere zwischen der 5. und der 7. Klasse.

4.2 Veränderungen der Mathematikkompetenzen

Die Entwicklung mathematischer Kompetenzen hat zwei Aspekte: Schüler werden sicherer im Lösen bereits bekannter Arten von Aufgaben, und sie lernen, neue Arten von Aufgaben zu lösen. Man kann annehmen, dass der Test T_2 für die Klasse 7 Aufgaben enthält, die Schüler der Klasse 5 noch nicht lösen können. Wie ich im nächsten Abschnitt zeige, folgen dann aus der Unterstel-

lung eines eindimensionalen Konstrukts erhebliche Fehler in der Erfassung der Kompetenzen.

Der genaue Inhalt der Items der Mathematiktests wurde zwar nicht veröffentlicht. Aus einer Publikation der Testentwickler (Neumann et al. 2013) kann man jedoch schließen, dass sich die Tests an der Idee orientieren, welche Mathematikkenntnisse Schüler einer bestimmten Klassenstufe haben sollten. Zum Beispiel heißt es bezüglich des Bereichs 'Data and chance':

> In Grade 5, children should be able to deal with data more systematically and purposefully than in kindergarten. Competence in this area is indicated by the extent to which children are able to collect data from simple experiments or observations and represent them in tables or figures such as bar charts or line charts. In the subarea 'chance' it is required to compare the probabilities of different events in random experiments and to know the basic concepts of 'certain', 'impossible', or 'likely'. Children should also be able to assess winning chances in dice games. (S. 90)

Für die Klassenstufe 9 wird dann gesagt:

> In Grade 9, students should be able to plan simple statistical studies, measure data systematically (e.g. distances covered by paper planes with different characteristics), organize data, and represent them graphically (e.g. by histograms or scatter plots). In order to analyze data, student of that age should be able to choose and apply suitable statistical methods (e.g. means or variance). This includes, for example, making conjectures on possible correlations between characteristics of a sample that are based on scatter plots. (S. 91)

Sicherlich wird also ein Test für die Klassenstufe 9 Aufgaben enthalten, die zu lösen Schüler der Klassenstufe 5 noch nicht gelernt haben. Man kann wohl annehmen, dass das auch für den Test in der Klassenstufe 7 (der nicht explizit kommentiert wird) zutrifft.

4.3 Fehler bei der Einschätzung von Veränderungen

Zu Beginn dieses Abschnitts wurde die Idee entwickelt, dass man sich für einen Vergleich der Ergebnisse der Tests T_1 und T_2 gedanklich auf einen Test T mit der zusammengefassten Itemmenge $J = J_1 \cup J_2$ beziehen sollte. Die leitende Frage lautet dann, wie sich die Kompetenzen der Schüler bezüglich dieses zusammengefassten Tests verändert haben. Man möchte wissen, wieviele korrekte

Antworten ein Schüler gegeben hätte, wenn dieser zusammenfassende Test in beiden Klassen durchgeführt worden wäre.

Offenbar benötigt man Annahmen darüber, wie die Schüler die jeweils nicht verwendeten Items aus J^b und J^a beantwortet hätten. Die durch (5) vorgenommene Gleichsetzung von Itemparametern impliziert solche Annahmen; ebenso sind jedoch andere Annahmen möglich, die zu wesentlich anderen Einschätzungen der Kompetenzentwicklung führen.

Um das genauer zu besprechen, verwende ich als ein formales Hilfsmittel das Konzept einer testcharakteristischen Funktion (im Folgenden kurz TC-Funktion). Bei einem Raschmodell beschreibt diese Funktion den Zusammenhang zwischen einem Theta-Wert und dem Erwartungswert für den Anteil der korrekt beantworteten Items. Die Definition für den Test T_k lautet:

$$g_k(\theta) = \frac{1}{m_k} \sum_{j \in J_k} \pi_{j,k}(\theta) = \frac{1}{m_k} \sum_{j \in J_k} \frac{\exp(\theta - \delta_{j,k})}{1 + \exp(\theta - \delta_{j,k})}$$

Dabei ist m_k die jeweilige Anzahl der Items ($m_1 = 23, m_2 = 22$); und $\pi_{j,k}(\theta)$ bezeichnet die Wahrscheinlichkeit einer richtigen Antwort beim Item j, wenn θ gegeben ist. Offenbar liefert diese TC-Funktion auch eine Möglichkeit, um aus einem beobachteten Summenscore s einen Schätzwert für Theta zu gewinnen:

$$\hat{\theta}_k(s) = \hat{g}_k^{-1}(s/m_k) \tag{6}$$

wobei $\hat{g}_k$ die TC-Funktion mit geschätzten Itemparametern bezeichnet. Im Folgenden verwende ich diesen Ansatz für einfache Illustrationen.

Die Methode, die beiden Tests dadurch vergleichbar zu machen, dass man die Parameter ihrer gemeinsamen Items gleichsetzt, impliziert eine Gleichsetzung der TC-Funktionen der beiden Tests, die somit auch als TC-Funktion für den zusammenfassenden Tests T angenommen wird. Abb. 3 zeigt diese TC-Funktion. Beispielsweise würde ein Schüler, der in beiden Tests 15 korrekte Antworten gegeben hat, jeweils einen Theta-Werte von etwa 1 erhalten.

Wie ich ausgeführt habe, kann die Annahme zeitinvarianter Itemparameter für den zusammenfassenden Test T nicht geprüft werden, und sie hat sehr unplausible Implikationen. Um zu illustrieren, welche Fehler bei der Erfassung von Kompetenzen resultieren können, nehme ich gleichwohl an, dass sich nicht nur die Parameter der gemeinsamen Items nicht verändern ($\delta_1^c = \delta_2^c$), sondern auch die der nicht wiederholten Items des ersten Tests ($\delta_1^a = \delta_2^a$). Dagegen

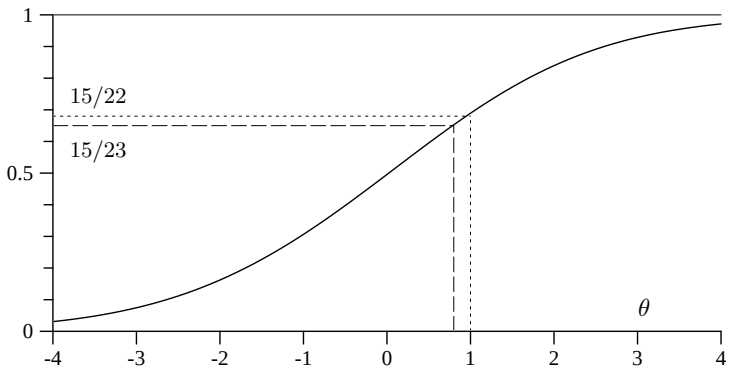

Abb. 3 Die aus einer Gleichsetzung von Itemparametern resultierende TC-Funktion für die Tests T_1, T_2 und T.

nehme ich jetzt an, dass die Schüler in der 5. Klasse noch nicht gelernt haben, die im zweiten Test neu hinzugekommenen Items zu lösen, also (wobei ich davon absehe, dass es sich teilweise um Multiple-Choice Items handelt): $\pi_{j,1} = 0$ für $j \in J^b$. Somit kann die in Abb. 3 gezeigte TC-Funktion zwar für den zusammenfassenden Test in der Klasse 7 verwendet werden, nicht jedoch in der Klasse 5.

Um einen zusammenfassenden Test für die Klasse 5 zu konstruieren, kann man folgendermaßen vorgehen. Angenommen, ein Schüler hat in dieser Klasse den Theta-Wert θ bezüglich T_1. Der Erwartungswert für den Anteil korrekt gelöster Aufgaben in diesem Test ist dann $g_1(\theta)$. Aus der Annahme, dass die Items in J^b noch nicht gelöst werden können, folgt für den Erwartungswert des Anteils korrekt gelöster Aufgaben bezüglich T: $g_1(\theta)\, m_1/m$, wobei $m = 39$ die Anzahl der Items in J ist. Verwendet man (6), erhält man als Theta-Wert bezüglich des zusammenfassenden Tests in der Klasse 5 den Wert

$$\theta^* = g^{-1}\big(\frac{m_1}{m}\, g_1(\theta)\big) \tag{7}$$

wobei g die TC-Funktion des zusammenfassenden Tests in der Klasse 7 bezeichnet.

Abb. 4 illustriert den Zusammenhang zwischen θ und θ^*. Zum Beispiel kann man sich auf einen Schüler beziehen, der aufgrund von T_1 in der Klasse 5 den Theta-Wert $\theta = 1$ erreicht hat. Dieser Schüler kann etwa 75 % der Items dieses Tests lösen. Das entspricht etwa 44 % der Items des zusammenfassenden

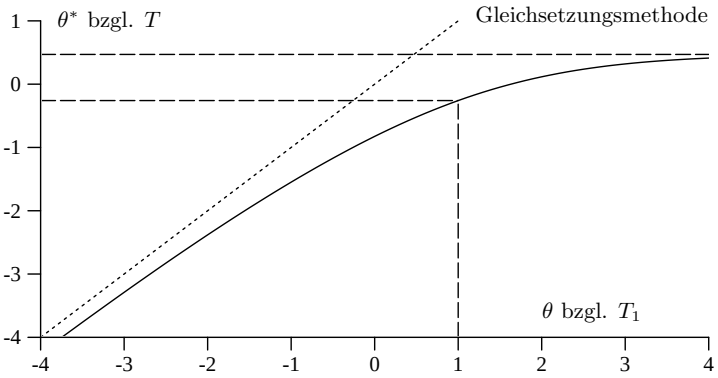

Abb. 4 Zusammenhang zwischen Theta-Werten in T_1 und durch (7) definierten Werten θ^* bzgl. des zusammenfassenden Tests in der Klasse 5.

Tests T, und bezüglich dieses Tests würde der Schüler also den Theta-Wert $\theta^* = -0.26$ erhalten. Die Annahme zeitinvarianter Itemparameter impliziert dagegen die gestrichelte Linie in der Abb. 4 und würde zu einem Theta-Wert $\theta^* = 1$ führen.

5 Vergleiche mit gemeinsamen Items

Die Überlegungen des vorangegangenen Abschnitts haben gezeigt, dass man aus den Ergebnissen der Tests T_1 und T_2 keine verlässlichen Informationen darüber gewinnt, wie sich die Kompetenzen bezüglich der Gesamtheit der verwendeten Items verändert haben. Wenn man sich für verlässlich einschätzbare Kompetenzveränderungen interessiert, kann man sich also nur auf die sechs in beiden Tests gemeinsam verwendeten Items beziehen. Im Folgenden möchte ich zeigen, dass man zur Quantifizierung wahlweise Summenscores oder Theta-Werte verwenden kann.

5.1 Summenscores und Theta-Werte

Verwendet man nur die sechs gleichen Items, kann man jedenfalls annehmen, dass beide Tests die gleiche Art mathematischer Kompetenz erfassen. Zur Quantifizierung ist eine Methode erforderlich, die ohne die Voraussetzung zeitinvarianter Itemparameter auskommt. Eine Möglichkeit besteht darin, Summenscores zu verwenden.

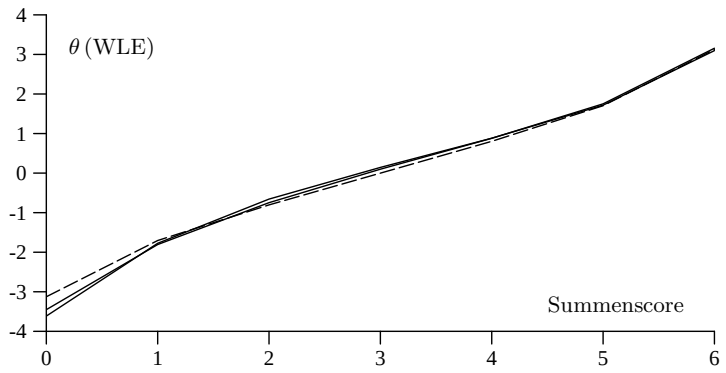

Abb. 5 Die durchgezogenen Linien zeigen den Zusammenhang zwischen Summenscores und Theta-Werten bei separat geschätzten Raschmodellen für die sechs gemeinsamen Items. Die gestrichelte Linie zeigt eine Skalentransformation aus gleichmäßig verteilten Itemparametern (s. Text).

Es ist natürlich möglich, Summenscores mit einem Raschmodell in Theta-Werte zu transfomieren. Man kann beispielsweise die Gleichung (6) verwenden. Eine etwas andere Methode der Skalentransformation, die auf sogenannten WLEs ('weighted likelihood estimates') beruht, wurde von Warm (1989) vorgeschlagen. Sie hat den Vorteil, dass man auch Theta-Werte für Schüler erhält, die alle oder überhaupt keine Items richtig gelöst haben.

Da Methoden der Skalentransformation von Itemparametern abhängen, stellt sich folgende Frage, wenn man für zwei Tests vergleichbare Theta-Werte konstruieren möchte: Ist es für die Skalentransformation erforderlich, identische Itemparameter zu verwenden? Jedenfalls in unserem Anwendungsfall ist das nicht erforderlich, denn wie Abb. 5 zeigt, erhält man auch unter Verwendung separat geschätzter Itemparameter praktisch identische Ergebnisse.

Nahezu identische Skalentransformationen sind also kein Hinweis auf nahezu identische Itemparameter. Tatsächlich kann man leicht Beispiele finden, bei denen vollständig unterschiedliche Itemparameter zu sehr ähnlichen Skalentransformationen führen. Zur Illustration zeigt die gestrichelte Linie in Abb. 5 eine Skalentransformation mit folgenden Parametern: $\tilde{\delta}_1 = -1.5$, $\tilde{\delta}_2 = -1.0$, $\tilde{\delta}_3 = -0.5$, $\tilde{\delta}_4 = 0.5$, $\tilde{\delta}_5 = 1.0$, $\tilde{\delta}_6 = 1.5$.

5.2 Zum Verständnis von Summenscores

Die Verwendung von Summenscores zur Quantifizierung von Kompetenzen kann mit dem Argument begründet werden, dass alle Items als gleichermaßen relevant – und in diesem Sinn: als austauschbar – gelten sollen. Dieses Argument impliziert, dass von sogenannten Itemschwierigkeiten abgesehen werden soll. Bei diesen Itemschwierigkeiten handelt es sich auch nicht um Eigenschaften der Items, sondern um Eigenschaften der Personen: Für eine Person, die ein Item lösen kann, ist es ein 'leichtes Item', und falls sie es nicht gelernt hat, ist es ein 'schwieriges Item'.

Wie die Itemparameter eines Raschmodells reflektieren auch Itemschwierigkeiten (wenn sie statistisch durch Proportionen nicht gelöster Items definiert werden) die Verteilung von Kompetenzen in einer für einen bestimmten Zeitpunkt definierten Population. Wenn man bei einer solchen Population Kompetenzen vergleichen möchte, kann es sinnvoll sein, eine bestimmte Verteilung der Kompetenzen (und somit auch der Itemschwierigkeiten) vorauszusetzen. Sobald man aber im Längsschnitt an Kompetenzveränderungen interessiert ist, muss man davon ausgehen, dass sich die Kompetenzverteilungen und somit auch die Itemschwierigkeiten verändern können. Man braucht also für sinnvolle Vergleiche ein Kompetenzmaß, das nicht von sich ändernden Itemschwierigkeiten abhängt.

Auch das Raschmodell setzt Austauschbarkeit der Items voraus, denn es impliziert, dass gleiche Summenscores zu gleichen Theta-Werten führen. Wenn ein Raschmodell gilt, kann man darin auch eine Begründung für Summenscores sehen; denn der Summenscore einer Person (bzw. der damit äquivalente Theta-Wert) genügt dann, um ihre itemspezifischen Lösungswahrscheinlichkeiten abzuleiten. Das Argument kann auch so formuliert werden: Wenn ein Raschmodell für eine gegebene Menge von Items gilt, spielt es keine Rolle, welche Items für einen Test ausgewählt werden. Dieses Argument setzt jedoch invariante Itemparameter voraus und liefert deshalb nur im Rahmen von Querschnittsanwendungen eine Begründung. Wenn man sich, wie in unserem Anwendungsfall, auf eine Itemmenge beziehen muss, die zwei Zeitpunkte umfasst, zwischen denen sich die Kompetenzverteilungen und die Itemparameter verändern, kann es kein Raschmodell geben, das beide Zeitpunkte umfasst.

Infolgedessen wird bei längsschnittlichen Anwendungen auch die Idee, dass

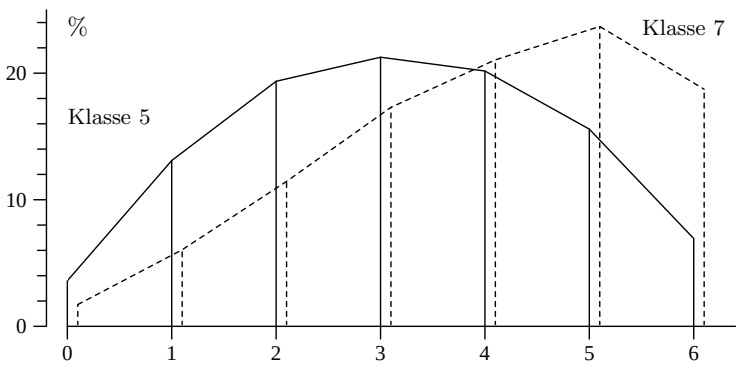

Abb. 6 Verteilungen der Summenscores bei den sechs in beiden Klassen gemeinsam verwendeten Items.

man mit einem Raschmodell Summenscores begründen kann, hinfällig. Deshalb verwendet auch die oben vorgeschlagene Begründung die Austauschbarkeit von Items in einer anderen Bedeutung. Es ist nicht gemeint, dass die Auswahl von Items für einen Test keine Rolle spielen sollte; sondern ganz im Gegenteil: Austauschbarkeit zwischen den Antworten auf die Items wird gefordert, weil die Bedeutung der zu erfassenden Kompetenz durch *alle* Items einer explizit bestimmten Menge von Items definiert wird.

5.3 Vergleiche mit Summenscores

Wenn man mit Summenscores Kompetenzen vergleicht, ergibt sich ihre Bedeutung aus den verwendeten Items. Abb. 6 zeigt einen Vergleich, der die sechs gemeinsamen Items verwendet. Offenbar haben sich die Schüler bezüglich einer durch diese Items definierten Kompetenz verbessert.

Aber wie in Abschnitt 4 ausgeführt wurde, erhält man daraus keine verläßlichen Informationen darüber, wie sich ihre Kompetenzen bezüglich aller in den beiden Tests verwendeten Items verändert haben. Ein solcher Vergleich hängt von Annahmen darüber ab, wie die Schüler die jeweils nicht verwendeten Items gelöst hätten. Nimmt man beispielsweise (wie in Abschnitt 4.3) an, dass die Schüler in der Klasse 5 noch nicht gelernt haben, die in der Klasse 7 neu hinzugekommenen Items zu lösen, ergäbe sich, wie Abb. 7 zeigt, ein vollständig anderer Vergleich.

Das gleiche Problem stellt sich, wenn man Kompetenzen zwischen unterschiedlichen Startkohorten vergleichen möchte. Als Beispiel vergleiche ich jetzt den Test der Mathematikkompetenz in der Klasse 7 mit einem Test der Mathematikkompetenz in der ersten Welle der Startkohorte Klasse 9.[6] Die beiden Tests haben wiederum sechs gemeinsame Items (nicht die gleichen, die zum Vergleich der Klassen 5 und 7 verwendet werden). Ich beziehe mich auf 6191 Schüler, die beim Test in der Klasse 7 mindestens eine gültige Antwort gegeben haben, und auf 14524 andere Schüler, die beim Test in der Klasse 9 mindestens eine gültige Antwort gegeben haben. Ein Likelihood-Ratio-Test der Annahme, dass man ein gemeinsames Raschmodell mit invarianten Itemparametern verwenden kann, liefert die Teststatisik 750.4 mit 5 Freiheitsgraden;[7] diese Annahme sollte also nicht gemacht werden. Da beide Tests dieselben sechs Items verwenden, ist dennoch ein sinnvoller Vergleich mit Summenscores möglich. Wie Abb. 8 zeigt, hat es einen Zuwachs in der *durch die sechs Items definierten* Kompetenz gegeben. Man kann jedoch sicher annehmen, dass ein Vergleich auf der Grundlage aller in den beiden Tests verwendeten Items zu anderen Ergebnissen führen würde.

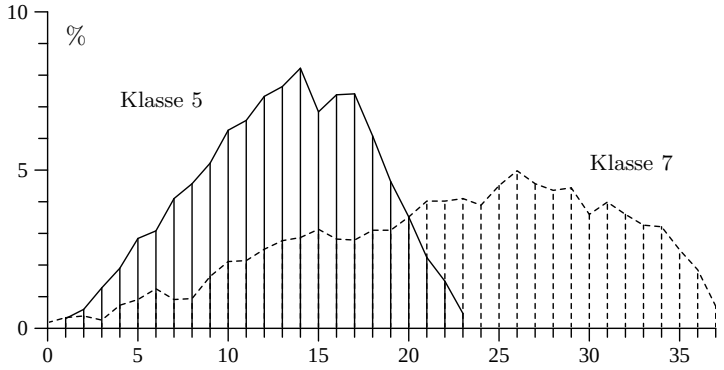

Abb. 7 Verteilungen der Summenscores bei einem Gesamttest T für beide Klassen unter der Annahme, dass Schüler der Klasse 5 noch nicht gelernt haben, die in der Klasse 7 neu hinzugekommenen Items zu lösen.

[6] Ich verwende die Version `doi:10.5157/NEPS:SC4:6.0.0`; siehe auch Fußnote 1.

[7] Für den Test wurden aus den 14524 Schülern der Klasse 9 9191 Schüler zufällig ausgewählt.

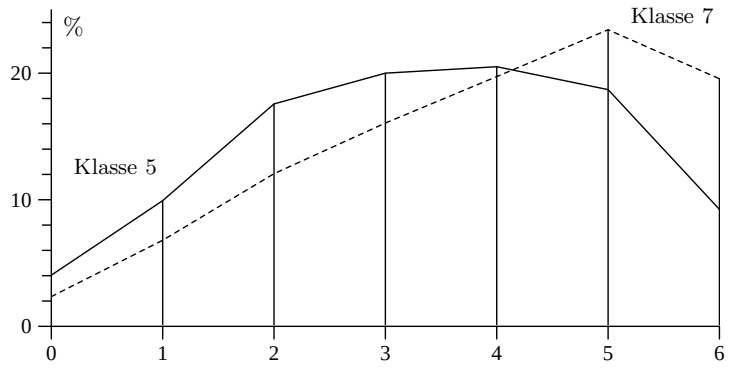

Abb. 8 Verteilungen der Summenscores bei sechs in den Klassen 7 und 9 gemeinsam verwendeten Items.

6 Zusammenfassung

Der Beitrag hat sich mit der Frage beschäftigt, wie man mit Daten aus zwei Tests, die im Rahmen des Nationalen Bildungspanels (NEPS) durchgeführt wurden, Veränderungen von Mathematikkompetenzen zwischen den Klassenstufen 5 und 7 erfassen kann. Es wird gezeigt, dass sich ein gemeinsames Raschmodell für die Testergebnisse hauptsächlich aus zwei Gründen nicht eignet. Erstens widersprechen die Testergebnisse der dafür erforderlichen Annahme zeitinvarianter Itemparameter. Zweitens ist ein solches Modell nicht mit Lernprozessen vereinbar, in denen Schüler lernen, neue Arten von Mathematikaufgaben zu lösen. Davon ist jedoch bei den hier verwendeten Mathematiktests auszugehen.

Um dennoch einen sinnvollen Vergleich zu ermöglichen, wird vorgeschlagen, sich auf einen zusammenfassenden Test zu beziehen, der alle Items umfasst, die in den beiden Klassenstufen eingesetzt worden sind. Dies verschafft eine Grundlage, auf der die tatsächlich erreichten Veränderungen der Mathematikkompetenz beurteilt werden können.

Um sich mit den verfügbaren Testergebnissen auf einen solchen zusammenfassenden Test zu beziehen, sind allerdings fragwürdige kontrafaktische Annahmen erforderlich, die zu erheblichen Fehlern bei der Einschätzung von Kompetenzveränderungen führen können. Wenn man an verlässlichen Einsichten

interessiert ist, kann man deshalb nur Items verwenden, die in beiden Tests identisch sind. Dann kann man sowohl Summenscores als auch Theta-Werte, die aus einer durch ein Raschmodell definierten Skalentransformation resultieren, verwenden, um Kompetenzveränderungen zu quantifizieren.

Literatur

Andersen, E.B. (1985). Estimating Latent Correlations Between Repeated Testings. *Psychometrika*, 50, 3–16.

Blossfeld, H.-P., Roßbach, H.-G, & von Maurice, J. (Eds.) (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, 14.

Duchhardt, C., & Gerdes, A. (2012). NEPS Technical Report for Mathematics – Scaling Results of Starting Cohort 3 in Fifth Grade. *NEPS Working Paper*, No. 17. Bamberg: NEPS.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). Linking the Data of the Competence Tests. *NEPS Survey Papers*, No. 1. Bamberg: LifBi.

Millsap, R. E. (2010). Testing Measurement Invariance Using Item Response Theory in Longitudinal Data: An Introduction. *Child Development Perspectives* 4, 5–9.

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and Assessing Mathematical Competence Over the Lifespan. *Journal for Educational Research Online* 5, 80–109.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement* 66, 63–84.

Stocking, M. L., & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement* 7, 201–210.

von Davier, M., & von Davier, A. (2007). A Unified Approach to IRT Scale Linking and Scale Transformations. *Methodology*, 3, 115–124.

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika* 54, 427–450.