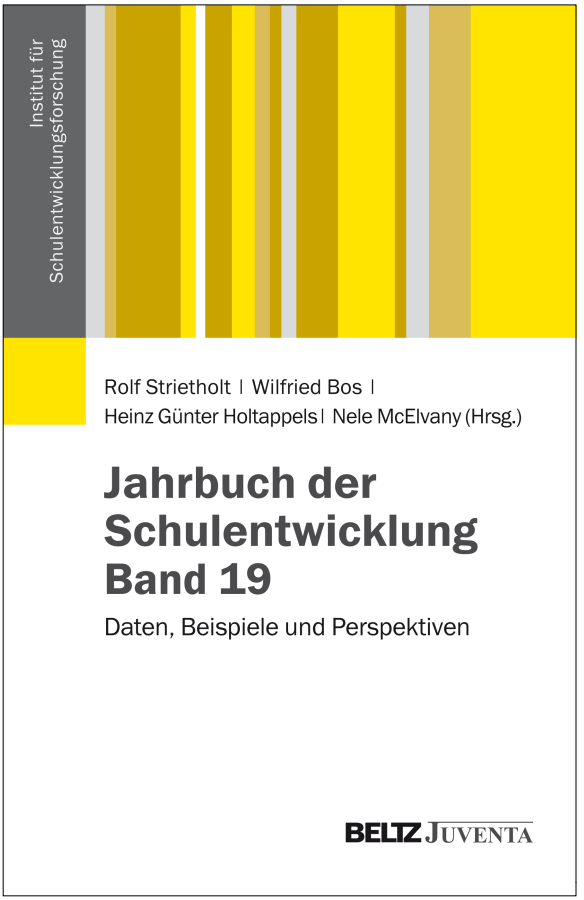


Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4.

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

I Konvergieren Leistungsprofile in Mathematik?

Evidenz aus fünf IEA Studien

Stefan Johansson & Rolf Strietholt

Internationale Schulleistungsstudien werden als neue globale Triebkraft im Bildungssektor diskutiert. Dabei wird argumentiert, dass sich Staaten an den am besten abschneidenden Bildungssystemen orientieren. Kritiker/-innen befürchten hier einen Verlust von Kreativität und Innovationen. Dabei gibt es bislang kaum empirische Evidenz für die internationale Homogenisierung von Leistungsprofilen. Mithilfe der Daten von fünf TIMSS Studien (1995–2011) widmet sich die vorliegende Studie der Frage nach der Konvergenz von Leistungsprofilen. Im Rahmen von latenten Profilanalyen betrachten wir die Leistungsprofile unterschiedlicher Staaten für vier mathematische Teildomänen (Algebra, Zahlen, Geometrie, Daten). Hierbei überprüfen wir die Hypothese, dass über die Zeit hinweg eine geringere Anzahl von Leistungsprofilen die unterschiedlichen Bildungssysteme beschreiben, d. h. Schüler/-innen weltweit zunehmend dieselben mathematischen Inhalte beherrschen. Die Analysen bestätigen, dass es möglich ist, latente Klassen für bestimmte Leistungsprofile zu identifizieren. Des Weiteren finden wir ähnliche Leistungsprofile in bestimmten Weltregionen und Staaten mit ähnlicher Sprache und Kultur. Die Analysen bieten allerdings keine Evidenz für eine Konvergenz von Leistungsprofilen.

Schlüsselwörter: Internationale Schulleistungsstudien, Mathematik, Vergleichende Erziehungswissenschaft, Weltcurriculum, TIMSS

1 Einführung

Die Rolle von Globalisierung für Bildung und Bildungssysteme ist ein kontroverses Thema. Die Globalisierung hat nicht nur ökonomisch gesehen zu einer zunehmenden internationalen Verflechtung geführt, sondern auch zu einem zunehmenden Wettbewerb von Bildungssystemen. So gilt die Aneignung von Wissen und Fähigkeiten als wichtige Bedingung für eine erfolgreiche individuelle Teilhabe an der Gesellschaft und das kognitive Leistungsniveau einer Gesellschaft gilt wiederum als wichtiger Prädiktor für Wirtschaftswachstum und Wettbewerbsfähigkeit in einer globalisierten Ökonomie (Hanushek & Woessmann, 2008; Sahlberg, 2006). Zusammen-

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

genommen führte das in den vergangenen Jahren zu einem gestiegenen Interesse an Bildungsergebnissen. Dieses gestiegene Interesse wird sichtbar an der zunehmenden Zahl von internationalen Schulleistungsstudien (TIMSS, PISA etc.), in denen Schülerleistung vergleichend erfasst wird. Diese Studien selbst werden vielfach als globale Triebkraft diskutiert, da sie von Medien breit rezipiert werden und deren Forschungsergebnisse eine prominente Position im politischen, professionellen und wissenschaftlichen Diskurs eingenommen haben (z. B. Baker & LeTendre, 2005; Gorur & Wu, 2014; Hegarty, 2014; Hopmann et al., 2007; Novóa & Yariv-Mashal, 2003; Simola, 2005). Des Weiteren scheinen die Forschungsergebnisse zunehmend bedeutsam für Steuerung und Reformen auf unterschiedlichen Ebenen des Bildungssystems (e.g., Grek, 2009, 2013; Ozga, 2012). In diesem Kontext wird argumentiert, dass die in internationalen Studien generierten Informationen für ein *policy borrowing* verwendet werden, d. h. dass sich Bildungssysteme an anderen als erfolgreich wahrgenommen Systemen orientieren, sodass internationale Schulleistungsstudien letztlich nationale Bildungssysteme beeinflussen. Als konkrete Beispiele für eine solche Steuerungsfunktion werden die Vereinheitlichung von Curricula und Unterrichtsinhalten, die Simplifizierung des Bildungsbegriffs und die Einführung einer neo-liberalen Agenda angebracht (Nordin & Sundberg, 2014). In jedem Fall wurden in den letzten Dekaden die Relevanz, Nutzung und Konsequenzen aus internationalen Schulleistungsstudien in einer Vielzahl von Publikationen diskutiert (z. B. Baker & LeTendre, 2005; Bos, 2002; Bos & Schwippert, 2003; Carvalho & Costa, 2014; Gustafsson, 2008; Pettersson, 2008; Spring, 2008; Strietholt et al., 2014).

1.1 Wechselwirkungen zwischen Bildungssystemen

In nationalen Bildungssystemen fordern unterschiedliche Akteure ständig Reformen, um wahrgenommenen Missständen zu begegnen („shop for repairs"; vgl. Baker & LeTentre, 2005, S. 151). In diesem Kontext bietet die Position im Leistungsranking einer internationalen Schulleistungsstudie ein empirisches Datum, um Missstände in einem Bildungssystem festzumachen. Bei der Suche nach Veränderungen und erfolgsversprechenden Reformen orientieren sich Staaten dann aneinander, um die wahrgenommenen Probleme zu beheben (e.g., Phillips & Ochs, 2003). Die Grundidee eines solchen *policy borrowing* ist, dass sich die schwächer abschneidenden Bildungssysteme an den Strategien von Bildungssystemen orientieren, die als erfolgreich wahrgenommen werden (wohingegen erfolgreiche Staaten Reformen implementieren, wo immer Raum ist). Damit erhalten Bildungssysteme, die als erfolgreich wahrgenommen werden, mehr Beachtung und

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

es kommt zu einer institutionellen Isomorphie, d. h. der Vereinheitlichung von Bildungssystemen. Ein potenzieller Nachteil einer solchen ist die institutionelle Isomorphie, in der die einzelnen Staaten ihre Einzigartigkeit und Kreativität verlieren und Innovationen ausbleiben (Pettersson, 2008; Wiseman, Astiz & Baker, 2014; Zhao, 2012). Offensichtliche Unterschiede zwischen den Staaten, wie etwa Sprache, Kultur oder Religion verlieren folglich an Bedeutung.

Es existieren unterschiedliche Erklärungsansätze und Theorien, wie eine globale Homogenisierung von Bildungssystemen abläuft. Meyer, Kamens und Benavot (1992) haben die Idee einer *Common World Educational Culture* (CWEC) entwickelt. In ähnlicher Weise beschreibt Dale (1999) eine *Globally Structured Agenda for Education* (GSAE). Beide Ansätze unterscheiden sich allerdings insofern, dass Dale argumentiert, dass sich Staaten eher regional, nicht aber global anpassen. Empirische Evidenz für beide Ansätze ist sehr überschaubar.

1.2 Lerngelegenheiten und Leistungsprofile im internationalen Vergleich

Es gibt unterschiedliche Ansätze um die Konvergenzhypothese zu untersuchen. Ein Anschlusspunkt bietet die Analyse der intendierten und implementierten Curricula in den einzelnen Bildungssystemen; dabei kann beispielsweise betrachtet werden, welche mathematischen Inhalte in nationalen Lehrplänen enthalten sind oder welche Lehrinhalte tatsächlich von den Lehrkräften unterrichtet werden. Baker und LeTendre (2005) haben hierfür Daten aus der Befragung der Lehrkräfte in TIMSS 1995 ausgewertet und herausgefunden, dass in allen Staaten eine Vielzahl von Inhaltsbereichen unterrichtet werden. Bemerkenswert war dabei insbesondere die Variation innerhalb von Bildungssystemen, da die Anzahl der behandelten mathematischen Inhaltsbereiche von Lehrkraft zu Lehrkraft deutlich variierte. Die Variation zwischen den Bildungssystemen war deutlich geringer, d. h. die am besten abschneidenden Staaten unterschieden sich nicht in Bezug auf das intendierte und implementierte Curriculum von den anderen Staaten. Unter Rückbezug auf die historische Bildungsforschung interpretieren Baker und LeTentre diesen Befund so, dass sich nationale Curricula im Zeitverlauf angeglichen haben. Dieser Argumentation folgend würde die zunehmende Anzahl an internationalen Schulleistungsstudien zu einer internationalen Homogenisierung der Lehrinhalte führen.

Ein anderer Ansatz zur Analyse der curricularen Konvergenz bieten die erreichten Kompetenzstände von Schüler/-innen. Wenn Lehrpläne und Unterrichtsinhalte international angeglichen werden, dann sollten sich die

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

Antwortmuster auf Testitems angleichen. So scheint es plausibel, dass ein verstärkter Fokus auf bestimmte Inhaltsbereiche (z. B. Geometrie) dazu führt, dass die Fähigkeit in diesem mathematischen Teilbereich höher ausgeprägt ist, als in anderen Teilbereichen. Der Vergleich der relativen Stärken und Schwächen bietet dabei einen Ansatz, um Unterschiede im Leistungsniveau der unterschiedlichen Staaten zu kontrollieren.

Relativ wenige Studien haben Fähigkeitsprofile unterschiedlicher Bildungssysteme betrachtet und im Zeitverlauf analysiert. Lie und Roe (2003) nutzten die Kompetenzdaten zum Lesen aus PISA 2000, um die Leistungsprofile der nordischen Staaten näher zu untersuchen. Dabei betrachteten sie die Anteile an korrekten Antworten zu den einzelnen Testitems. Diese Analysen zeigten, dass Dänemark, Norwegen und Schweden ähnliche Kompetenzprofile haben, d. h. in den Leistungstest bei denselben Items Stärken bzw. Schwächen aufweisen. Demgegenüber wies Finnland ein anderes Kompetenzprofil auf. Die Autoren vermuten, dass Schüler/-innen aus Finnland unter anderem besser in der Lage sind, zwischen den Zeilen zu lesen, was ein Grund für das gute Abschneiden Finnlands in PISA ist. Da die Studie ausschließlich Daten aus dem ersten PISA-Zyklus nutzt, liegen keine Informationen über mögliche Trends vor.

Mit einem ähnlichen Analyseansatz haben Kjarnsli und Lie (2008) TIMSS 2003 Daten ausgewertet, um die relativen Stärken und Schwächen in Naturwissenschaften in einer größeren Stichprobe von Staaten zu untersuchen. Sie nutzten die Residuen der Anteile korrekter Antworten für die Items des naturwissenschaftlichen Tests im Rahmen von Clusteranalysen, um Staaten mit ähnlichen Profilen zu identifizieren. Die Cluster wurden in Bezug auf räumliche und sprachlich-kulturelle Regionen beschrieben (z. B. Ostasien, Südosteuropa, englischsprachig, Arabisch). In weiteren Analysen haben die Autoren Zusammenhänge zwischen der Clusterzugehörigkeit und anderen Systemmerkmalen untersucht, wobei insbesondere die Staaten in dem englischsprachigen Cluster insgesamt besser in den Leistungstests abschnitten und offene Testaufgaben (constructed response) besser beherrschten. Demgegenüber hatten die südeuropäischen Staaten eine relative Stärke bei den Multiple-Choice-Aufgaben. Wenngleich die Cluster augenscheinlich mit Sprache und Kultur zusammenhängen, so schließen Kjarnsli und Lie, dass ein linguistischer Faktor nicht allein für die Clusterung der Staaten verantwortlich ist.

Die querschnittliche Identifikation bestimmter Leistungsprofile sagt noch nichts über Homogenisierungsprozesse aus: Beschreiben weniger Profile im Zeitverlauf mehr Staaten? Zur Beantwortung dieser Frage haben Rutkowski und Rutkowski (2009) die Daten der TIMSS-Zyklen aus den

Jahren 1995, 1999 und 2003 analysiert. Die leitende Forschungsfrage war dabei die empirische Prüfung der Theorien zur weltweiten Homogenisierung von Leistungsprofilen in Mathematik (Meyer et al., 1997; Dale, 1999). In diesem Studiendesign bilden Staaten (n = 16), die wiederholt beobachtet werden, die Analyseebene. Untersucht werden die Anteile an korrekt gelösten Mathematikaufgaben, wobei wiederum Residuen (relative Stärken und Schwächen) betrachtet werden. Mithilfe von Clusteranalysen identifizierten Rutkowski und Rutkowski ähnliche Leistungsprofile, die mit räumlichen und sprachlich-kulturellen Regionen (z. B. Ostasien, Südosteuropa, englischsprachig) korrespondieren. Im Zeitverlauf erweist sich die Zuordnung der Staaten allerdings als stabil, das heißt, dass die Analysen keine empirische Evidenz für eine curriculare Konvergenz im Zeitverlauf bieten. Die Studie stellt folglich auch keine Evidenz dafür dar, dass internationale Schulleistungsstudien Curricula homogenisieren. Gleichzeitig muss darauf hingewiesen werden, dass der Zeitraum, den die Studie abdeckt, zu kurz sein mag, um globale Trends zu identifizieren und Evidenz für die Homogenisierungshypothese zu bieten. Die Einführung des Internets, einfacheres weltweites Reisen und internationaler Handel sind heutzutage weit fortgeschritten; gleichzeitig handelt es sich bei diesen und anderen Bereichen der Globalisierung um Entwicklungen, die besonders deutlich aus einer längerfristigen Perspektive sichtbar werden. Die Betrachtung der Residuen einzelner Testitems markiert eine weitere Einschränkung der besprochenen Studien. Curricula mögen sich in einzelnen Items manifestieren, gleichzeitig ist die Generalisierbarkeit einzelner Items gering und der Messfehler hoch. Anstelle von Einzelitems sind inhaltliche Teilbereiche einer Leistungsdomäne (z. B. Geometrie als Teil der Mathematik) vermutlich eine geeignetere Analyseebene für globale Konvergenzen in Leistungsergebnissen.

1.3 Die vorliegende Studie

Vor dem Hintergrund der zuvor diskutierten Einschränkungen früherer Forschung untersucht die vorliegende Studie die curriculare Konvergenz über einen längeren Zeitraum und auf Ebene von mathematischen Teilkompetenzen. Hierbei greifen wir auf insgesamt fünf TIMSS-Zyklen zurück, um einen Zeitraum von 16 Jahren abzudecken. Wir nehmen an, dass mehrere Jahre vergehen, ehe sich Reformen in Curricula tatsächlich in der Schülerleistung niederschlagen, sodass es notwendig ist, einen längeren Zeitraum zu betrachten. Des Weiteren argumentieren wir, dass Curricula nicht auf Einzelitemebene definiert werden sollten, sondern auf einem höheren Abstraktionsniveau. Daher betrachten wir Fähigkeitsprofile auf Ebene von fachlichen Teildomänen (hier: mathematische Teilbereiche), um

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

curriculare Konvergenz empirisch zu untersuchen. Das Hauptziel der Studie ist dabei die längsschnittliche Analyse von Leistungsprofilen in unterschiedlichen Staaten. Wenn sich die Leistungsprofile globalen (CWEC Hypothese) oder regionalen (GSAE Hypothese) Trends angleichen, ist das ein Indikator für eine Harmonisierung von Bildungssystemen bzw. deren Curricula. Eine solche Konvergenz wäre ein Hinweis auf Globalisierungskräfte; eine mögliche Erklärung wäre dann beispielsweise die gestiegene Verbreitung internationaler Schulleistungsstudien.

2 Methode

2.1 Stichprobe

In der vorliegenden Studie nutzen wir die Daten der Sekundarstufe (IEA Population B) aus der *Third International Mathematics and Science Study* (TIMSS). In einigen Staaten wurden mehrere Jahrgangsstufen getestet (z. B. 7. und 8. Jahrgangsstufe); in diesem Fall haben wir die Daten aus der Jahrgangsstufe verwendet, zu der für die meisten Jahre Daten vorliegen. So werden im Zeitverlauf innerhalb der Staaten möglichst homogene Gruppen von Schüler/-innen verglichen. Die Gesamtstichprobe besteht aus 1.066.417 Schüler/-innen aus 87 Staaten. Die Daten wurden in den Jahren 1995 ($n_{Schüler/-innen}$ = 136.973, $n_{Staaten}$ = 40), 1999 ($n_{Schüler/-innen}$ = 237.833, $n_{Staaten}$ = 35), 2003 ($n_{Schüler/-innen}$ = 237.833, $n_{Staaten}$ = 51), 2007 ($n_{Schüler/-innen}$ = 245.553, $n_{Staaten}$ = 57) und 2011 ($n_{Schüler/-innen}$ = 281.995, $n_{Staaten}$ = 49) erhoben. 15 Bildungssysteme haben an allen fünf, 10 Staaten an vier, 21 Staaten an drei, 13 Staaten an zwei und 28 Staaten an nur einer Erhebung teilgenommen.

2.2 Instrumente

Die in TIMSS eingesetzten Leistungstests haben ein sogenanntes Multi-Matrix-Design, bei dem die Schüler/-innen jeweils nur einen Teil der Testaufgaben bearbeiten. Tab. 1 zeigt, dass die Zahl der Testaufgaben im Lauf der Zeit kontinuierlich gestiegen ist, wobei parallel auch der Anteil an Items mit offenem Aufgabenformat (constructed response) im Vergleich zu den Multiple-Choice-Aufgaben kontinuierlich erhöht wurde. Der Überblick zeigt auch, dass die Tests in allen Studien zweimal 45 Minuten gedauert haben. Ein Teil der Aufgaben von früheren Erhebungszyklen wurde in die Tests neuer Erhebungszyklen integriert, um die Tests zu verlinken und Schülerleistung auf derselben Metrik abbilden zu können (vgl. Martin & Mullis, 2012).

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

Tabelle 1: Beschreibung der Leistungstests der unterschiedlichen TIMSS-Zyklen

	TIMSS95	TIMSS99	TIMSS03	TIMSS07	TIMSS11
Itemanzahl	151 (125 MC, 26 CR)	162 (125 MC, 37 CR)	194 (128 MC, 66CR)	215 (117 MC, 98 CR)	217 (118 MC, 99 CR)
Testzeit	45 + 45 min	45 + 45 min	45 + 45 min	45 + 45 min	45 + 45 min
Mathematische Inhaltsbereiche	Fractions & Number Sense (51) Measurement (18)	Fractions & Number Sense (61) Measurement (24)	Number (57) Measurement (31)	Number (63)	Number (61)
	Algebra (27)	Algebra (35)	Algebra (47)	Algebra (64)	Algebra (70)
	Geometry (23) Proportionality (11)	Geometry (21)	Geometry (31)	Geometry (47)	Geometry (43)
	Data Representation, Analysis, & Probability (21)	Measurement (24) Data Representation, Analysis, & Probability (21)	Data (28)	Data & Chance (41)	Data & Chance (43)

Anmerkung: MC = Multiple Choice, CR = Constructed Response.

Mathematik kann in unterschiedliche Inhaltsbereiche gegliedert werden, die zusammengenommen das allgemeine Konstrukt ‚Mathematik' definieren. Auf welche Teilbereiche dabei zurückgegriffen wird, ist abhängig von der Forschungsfrage, beispielweise vom Alter der Schüler/-innen. Tab. 1 listet die mathematischen Teilbereiche auf, die in den einzelnen Erhebungswellen separat ausgewiesen werden. Die Übersicht zeigt augenscheinliche Veränderungen; so wird in den jüngeren Erhebungswellen zwischen den vier mathematischen Teildomänen *Number*, *Algebra*, *Geometry* und *Data & Chance* unterschieden, wohingegen bei den älteren Studien eine andere Unterteilung vorgenommen wurde. Gleichzeitig ist offensichtlich, dass es sich nicht um inhaltliche Unterschiede handelt, denn eine Reihe von Differenzen beruht nur auf terminologischen Veränderungen. In den älteren Studien wurde beispielsweise der Betriff *Fractions and Number Sense* verwendet und in den jüngeren Studien der Begriff *Number*. Die verbliebenen Unterschiede sind auf einen höheren Differenzierungsgrad in den jüngeren Studien zurückzuführen. So wurden 1995 *Measurement* und *Proportionality* als separate Teildimensionen ausgewiesen, wohingegen diese Bereiche in den jüngeren Studien Teil der Inhaltsbereiche *Number* bzw. *Geometry* sind.

2.3 Analysestrategie

Anknüpfend an diese Unterscheidung zwischen den mathematischen Teilbereichen *Number*, *Algebra*, *Geometry* und *Data & Chance* betrachten wir die Leistungsprofile der unterschiedlichen Bildungssysteme, indem wir die Testscores der Schüler/-innen in den vier Inhaltsbereichen für die Staaten in den unterschiedlichen Jahren aggregieren. Hierzu führen wir latente

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

Profilanalysen durch. Basierend auf den latenten Klassen für unterschiedliche Leistungsprofile untersuchen wir in einem weiteren Analyseschritt die Hypothese einer globalen Konvergenz von Curricula; hierzu vergleichen wir die Klassenzugehörigkeit der einzelnen Staaten über die unterschiedlichen Erhebungswellen von TIMSS hinweg. Dieses Vorgehen beruht auf der Annahme, dass sich Veränderungen in Curricula und Lehrplänen letztendlich in den Fähigkeiten in den jeweiligen mathematischen Teildomänen niederschlagen.

2.3.1 Gemeinsame Skalierung der Leistungsdaten

Die Leistungsscores für die mathematischen Teildomänen sind nicht über die unterschiedlichen Erhebungswellen hinweg vergleichbar. Daher skalieren wir in einem ersten Schritt die Leistungstests im Rahmen eines *common-item nonequivalent group design* neu; hierbei erscheint zwar kein Item in allen Erhebungswellen, aber ein Teil der Aufgaben erscheint in mehreren Wellen (Kolen & Brennan, 2004; Strietholt & Rosén, 2016). Die Skalierung erfolgt separat für die vier Inhaltsbereiche (s. Tab. 2 für die Anzahl der Items in den einzelnen Wellen). Für den Bereich *Number* liegen insgesamt 227 Items, für *Algebra* insgesamt 150 Items, für *Geometry* insgesamt 128 (inklusive der Items aus dem Bereich *Measurement*) und für *Data & Chance* insgesamt 92 Items vor (inklusive der Items aus dem Bereich *Proportionality*).

Die Verlinkung erfolgt mithilfe von Modellen der Item Response Theorie (IRT) und im Rahmen einer gleichzeitigen Kalibrierung aller Itemparameter (Kim & Cohen, 1998, 2002). Für Multiple-Choice-Items verwenden wir das dreiparametrische logistische (3PL), für binäre (richtig/falsch) Constructed-Response-Items das zweiparametrische logistische (2PL) und für Constructed-Response-Items, bei denen Schüler/-innen bis zu drei Teilpunkte erhalten können, das Generalisierte Partial Credit Modell. Das R-Paket des Test Analysis Moduls (TAM; Kiefer, Robitzsch & Wu, 2015) wurde für die Mehrgruppen-IRT-Analysen verwendet. Für die unterschiedlichen Teilstichproben (Staaten*Jahre) wurden normalverteilte Fähigkeitsparameter angenommen. Aufgrund der großen Gesamtstichproben wurde für die Itemparameterschätzungen aus jeder Teilstichprobe eine Zufallsstichprobe von 500 Schüler/-innen gezogen.

Basierend auf den Itemparametern wurden dann in einem separaten Schritt für alle Schüler/-innen der Gesamtstichprobe EAP (expected a posteriori) Personenparameter geschätzt. Die Personenparameter wurden auf eine Metrik mit einem internationalen Mittelwert von 500 und einer Standardabweichung von 100 transformiert. Die standardisierten Leis-

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

tungsscores für die Bereiche *Number*, *Algebra*, *Geometry* und *Data & Chance* wurden anschließend – unter Berücksichtigung der Stichprobengewichte (houwgt) – auf Staatenebene aggregiert. Hieraus resultierte ein Datensatz mit 232 Fällen, wobei die Fälle Staaten sind, die an bis zu fünf TIMSS-Zyklen teilgenommen haben.

Im letzten Schritt der Datenvorbereitung wurden Residualscores für die Teilbereiche *Number*, *Algebra*, *Geometry* und *Data & Chance* berechnet, um Aussagen darüber zu treffen, wie gut ein Bildungssystem in einem mathematischen Teilbereich relativ zu den anderen Teilbereichen abschneidet. Die Residualscores wurden gebildet, indem zunächst für jede Teilstichprobe der Mittelwert über die vier Teildomänen berechnet wurde, um diesen Gesamtscore dann von den jeweiligen Scores in den vier Teildomänen zu subtrahieren. Die Residualscores für die vier Teildomänen geben über relative Stärken bzw. Schwächen der jeweiligen Teilstichproben Auskunft.

2.3.2 Latente Profilanalysen

Die Hypothese dieser Arbeit ist, dass es Staaten mit ähnlichen mathematischen Leistungsprofilen gibt und die Staaten sich im Zeitverlauf weiter angleichen. Zur empirischen Prüfung dieser Hypothese führen wir latente Profilanalysen (LPA) durch, wobei die Staaten anhand der Residualscores für *Number*, *Algebra*, *Geometry* und *Data & Chance* klassifiziert werden. Zunächst vergleichen wir hierzu eine Reihe von Modellen anhand unterschiedlicher Evaluationskriterien, um die Anzahl der latenten Klassen zu bestimmen.

Das *Bayesian Information Criterion* (BIC; Schwarz, 1978) ist ein Likelihood-basiertes Kriterium, das Modellkomplexität – d. h. die Anzahl an Parametern im Modell – bestraft. Ein geringerer Wert beim BIC steht für ein besseres Modell. Zusätzlich zu diesem statistischen Kriterium betrachten wir den Wert und den Nutzen der Klassifikation anhand der Entropy. Entropy ist ein Maß für die Trennbarkeit der latenten Klassen und die Präzision, mit der die Fälle den einzelnen Klassen zugewiesen werden. Dabei kann die Entropy Werte zwischen 0 und 1 annehmen, wobei hohe Werte für eine klare Trennbarkeit der Zuordnung der Fälle zu den Klassen stehen. Des Weiteren betrachten wir die Größe der einzelnen Klassen, wobei eine übermäßig große Anzahl von Klassen problematisch sein kann, sofern einzelne Klassen nur eine kleine Zahl von Fällen repräsentieren. Hierdurch ist die Komplexitätsreduktion gering und die Gefahr statistischer Artefakte steigt (Samuelsen & Dayton, 2010). Letztlich betrachten wir bei den unterschiedlichen Modellen die Profile in den einzelnen Klassen. Das Augenmerk liegt dabei darauf, ob bei zunehmend komplexeren Modellen distinkte

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

Klassen extrahiert werden. Gut interpretierbar sind qualitative Unterschiede bei den einzelnen Profilen. Demgegenüber sprechen parallele Profile eher für graduelle Unterschiede innerhalb eines übergeordneten Profils, das gilt insbesondere für den Fall, wenn die parallelen Profile eng zusammen liegen.

Bei der Modellspezifikation ist zu beachten, dass die Residualscores für *Number*, *Algebra*, *Geometry* und *Data & Chance* konstruktionsbasiert nicht unabhängig voneinander sind. Durch die Standardisierung (s. o.) ist die Summe der Residualscores für jeden Fall gleich 0. Daher werden innerhalb aller Klassen die Mittelwerte (M) eines beliebigen Residualscores (hier: *Data & Chance*) folgendermaßen fixiert: $M_{Data\ \&\ Chance} = -(M_{Number} + M_{Algebra} + M_{Geometry})$. Des Weiteren werden in allen Klassen die Varianzen der Residualscores frei geschätzt, wohingegen die Varianzen derselben Residualscores über alle Klassen hinweg gleichgesetzt werden.

3 Ergebnisse

3.1 Leistungsprofile in Mathematik

Zur Beantwortung der Konvergenzhypothese untersuchen wir die Leistungsprofile bei den 232 Fällen. Hierzu schätzen wir in einem ersten Schritt eine Reihe von LPA-Modellen, um explorativ die Anzahl der Leistungsprofile zu bestimmen, anhand derer die Fälle klassifiziert werden. Grundlage bilden dabei die Residualscores der vier mathematischen Teilkompetenzen. Tab. 2 fasst die Fitkriterien für Modelle mit 1 bis 8 latenten Klassen zusammen. Der Vergleich zeigt, dass sich BIC und Entropy durch die Hinzunahme weiterer latenter Profile fast durchgängig verbessern, wobei die Verbesserung beider Fitkriterien nicht sprunghaft, sondern graduell verläuft.

Tabelle 2: Fitkriterien für die LPA Modelle

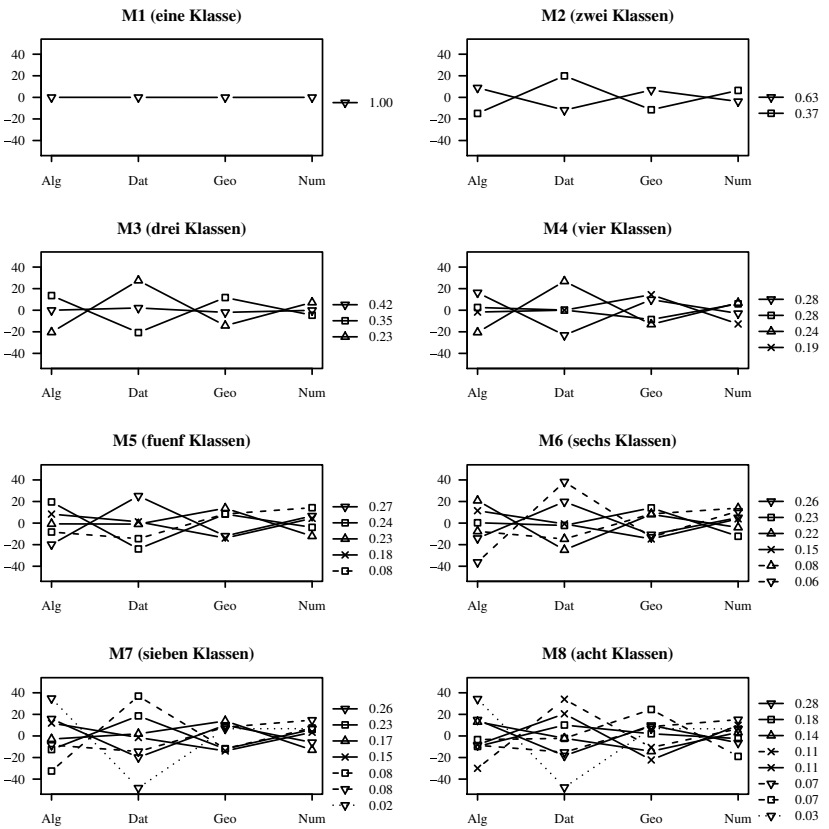
Modell	M1	M2	M3	M4	M5	M6	M7	M8
Anzahl Klassen	1	2	3	4	5	6	7	8
Anzahl Parameter	7	11	15	19	23	27	31	35
BIC	7809	7617	7585	7535	7507	7480	7455	7434
Entropy	1	0,77	0,76	0,77	0,82	0,86	0,87	0,89
Anteil an Fällen in den einzelnen Klassen	100%	63%	42%	28%	27%	26%	26%	28%
		37%	35%	28%	24%	23%	23%	18%
			23%	24%	23%	22%	17%	14%
				19%	18%	15%	15%	11%
					8%	8%	8%	11%
						6%	8%	7%
							2%	7%
								3%

Anmerkung. Der Anteil an Fällen in den einzelnen Klassen basiert auf den geschätzten Modellparametern

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

Die einzelnen Klassen innerhalb der Modelle repräsentieren unterschiedlich viele Fälle. In Bezug auf die Klassengrößen fällt auf, dass bei den Modellen M1 bis M4 selbst die kleinsten Klassen noch einen substantiellen Anteil der Fälle repräsentieren, wohingegen die jeweils kleinsten Klassen bei den komplexeren Modellen (M5 bis M8) nur noch für weniger als zehn Prozent der Fälle stehen. Würde man einen Anteil von zehn Prozent als Schwellenwert für die Extraktion von Klassen anlegen, so spricht dieses Kriterium gegen die Modelle M5 bis M8.

Abbildung 1: Leistungsprofile in mathematischen Teildomänen für die LPA Modelle



Anmerkung. In den jeweiligen Legenden ist der Anteil an Fällen in den einzelnen Klassen angegeben.

Um zu evaluieren, ob die einzelnen Klassen auch distinkte Profile repräsentieren, betrachten wir im Weiteren die einzelnen Profile aller Modelle (vgl. Abb. 1). Beispielsweise besteht Modell M1 aus nur einer Klasse, wobei, be-

dingt durch die Standardisierung der Residualscores, alle Mittelwerte für *Number*, *Algebra*, *Geometry* und *Data & Chance* gleich 0 sind. Bei Modell M2 entstehen zwei distinkte Leistungsprofile: Das eine Profil zeigt relative Stärken in den Bereichen *Algebra* und *Geometry* und relative Schwächen in den Bereichen *Number* und *Data & Chance*, bei dem anderen Profil sind die relativen Stärken und Schwächen gegensätzlich ausgeprägt. Wir interpretieren die Profile der Modelle M1 bis M4 so, dass durch die Hinzunahme weiterer Klassen distinkte Profile entstehen. Demgegenüber entstehen bei den Modellen M5 bis M8 durch die Hinzunahme neuer Klassen keine neuen distinkten Profile, sondern Profile, die graduelle Unterschiede bereits zuvor extrahierter Profile darstellen. Zur Veranschaulichung haben wir in der Abbildung distinkte Profile mithilfe unterschiedlicher Symbole markiert und graduelle Unterscheide durch unterschiedliche Linientypen. Offensichtlich ist eine eindeutige Unterscheidung zwischen distinkten und graduellen Profilunterschieden anhand der Plots nicht immer möglich. Gleichzeitig ist festzustellen, dass die Profile, die unseres Erachtens graduelle Unterschiede beschrieben, jeweils nur einen geringen Anteil an Fällen repräsentieren.

Zusammenfassend legen die Evaluationskriterien BIC und Entropy zwar die Extraktion von mehr Klassen nahe, gleichzeitig erscheint uns der Erkenntniszugewinn bei einer Extraktion von mehr als vier Klassen gering, da keine neuen distinkten Klassen extrahiert werden und die neuen Klassen auch nur einen geringen Anteil von Fällen beschreiben. Zudem ist die Entropy ist bei allen Modellen größer als .75; d. h. die Zuweisungsgenauigkeit der Fälle zu den einzelnen Profilen ist bei allen Modellen in einem akzeptablen Bereich. Vor diesem Hintergrund entscheiden wir uns – lex parsimuniae – für die weiteren Analysen für das Modell mit vier Klassen.

Die vier latenten Leistungsprofile aus Modell M4 repräsentieren zwischen 19 und 29 Prozent der Fälle. Während zwei Klassen relativ ausgeglichene Leistungsprofile aufweisen, zeichnen sich die beiden anderen Klassen durch extremere Unterschiede in den einzelnen mathematischen Teilkompetenzen aus. Die beiden ausgeglichenen Profile unterscheiden sich hauptsächlich bei den Teilkompetenzen *Geometry* und *Number*, wobei die eine Klasse relativ stark in *Geometry* und relativ schwach in *Number* ist (19% der Fälle), wohingegen in der anderen Klasse relative Stärken und Schwächen diametral ausfallen (28%). Die beiden Profile mit den weniger ausgeglichenen Profilen unterscheiden sich am deutlichsten in den Bereichen *Algebra* und *Data & Chance*, wobei das eine Leistungsprofil ausgesprochen schwach in *Algebra* und ausgesprochen stark in *Data & Chance* ist (24%) und das

andere Leistungsprofil ausgesprochen stark in *Algebra* und ausgesprochen schwach in *Data & Chance* ist (28%).

Wir haben nun Leistungsprofile identifiziert, anhand derer Staaten in Bezug auf ihre relative Schwächen und Stärken in mathematischen Teildomänen klassifiziert werden können. Als nächstes weisen wir die unterschiedlichen Bildungssysteme den einzelnen latenten Klassen zu, um die Konvergenzhypothese zu überprüfen.

3.1 Konvergenz im Zeitverlauf

Wie verändern sich die Leistungsprofile unterschiedlicher Bildungssysteme über die Zeit hinweg? Um die Konvergenzhypothese zu untersuchen, betrachten wir, ob sich im Zeitverlauf mehr Staaten durch weniger Leistungsprofile beschreiben lassen. Da wir Veränderungen untersuchen, betrachten wir ausschließlich die 57 Staaten, die zu mindestens zwei Zeitpunkten an TIMSS teilgenommen haben. In Abb. 2 sind die Leistungsprofile dieser Staaten für die unterschiedlichen Messzeitpunkte abgebildet. Dabei stehen die vier Symbole für die in Modell M4 identifizierten Leistungsprofile. Um das Interpretieren der Ergebnisse zu vereinfachen, wurden die Staaten nicht alphabethisch, sondern bezüglich der Klassifizierungsmuster geordnet.

Wir untersuchen die Hypothese, dass Leistungsprofile im Zeitverlauf konvergieren. In Bezug auf globale Konvergenz bedeutet das, dass in Tab. 3 einzelne Symbole in den jüngeren Studienzyklen dominanter werden, während gleichzeitig andere Symbole weniger häufig erscheinen. Dieses Muster bestätigt sich allerdings nicht; vielmehr gibt es innerhalb vieler Staaten eine bemerkenswerte Stabilität der relativen Stärken und Schwächen im Zeitverlauf. Daneben gibt es zwar eine Reihe von Staaten, bei denen im Zeitverlauf Veränderungen bei den Profilzuordnungen zu finden sind, es ist aber international keine globale Konvergenz zu einem bestimmten Leistungsprofil erkennbar. Die Anteile der einzelnen Klassen variieren zwar etwas im Teilverlauf (vgl. Tab. 3), die Konfidenzintervalle legen allerdings nahe, dass geringe Schwankungen zwischen einzelnen Jahren nicht überinterpretiert werden sollten. Wichtiger ist vielmehr festzuhalten, dass die Daten keine Evidenz für einen Trend bieten, dass ein Profil im Zeitverlauf dominanter wird.

Ein genauerer Blick auf die Leistungsprofile der einzelnen Staaten legt regionale, kulturelle und sprachliche Ähnlichkeiten nahe. Hierbei lassen sich unterschiedliche Cluster von Bildungssystemen identifizieren. Im oberen Teil von Tab. 3 finden sich die Staaten, die in 2011 ein Leistungsprofil aufweisen, das in den Bereichen *Data & Chance* und *Number* durch relative Stärken und in den Bereichen *Algebra* und *Geometry* durch relative Schwä-

Leseprobe aus: Strietholt/Bos/Holtappels/McElvany (Hrsg.), Jahrbuch der Schulentwicklung, Band 19
ISBN 978-3-7799-0919-4 © 2016 Beltz Verlag, Weinheim Basel
http://www.beltz.de/de/nc/verlagsgruppe-beltz/gesamtprogramm.html?isbn=978-3-7799-0919-4

chen gekennzeichnet ist (markiert mit dem Dreieck, Spitze oben). Dieses Leistungsprofil findet sich über den gesamten Zeitraum in den skandinavischen Staaten Norwegen und Schweden sowie einer Reihe von anglophonen europäischen Staaten des damaligen British Empires (England, Neuseeland, Schottland). Dieser Befund verweist auf regionale, kulturelle und sprachliche Ähnlichkeiten. In Bezug auf Veränderungen scheint der Einfluss des British Empires noch heute bedeutsam zu sein; so wechselten Australien, Kanada und die USA das Leistungsprofil, sodass hier auch ein Hinweis auf eine sprachliche und kulturelle curriculare Konvergenz vorliegt.

Im unteren Teil der Tabelle sind die Staaten aufgeführt, die – im Gegensatz zu dem zuvor beschriebenen Staatencluster – in den Bereichen *Data & Chance* und *Number* relative Schwächen und in *Algebra* und *Geometry* relative Stärken haben (markiert mit dem Dreieck, Spitze unten). Auch hier finden sich mit Russland, Rumänien, Mazedonien, Serbien und Montenegro, Moldawien, Bulgarien, Armenien und Georgien eine Reihe von Staaten, die offensichtliche Ähnlichkeiten aufweisen. Die genannten Bildungssysteme befinden sich in einer geographischen Region und sind geschichtlich eng miteinander verknüpft (z. B. postsowjetische Staaten, Warschauer Pakt, politisch-ideologisch).

Während die beschriebenen Cluster eine gewisse Rolle regionaler, kultureller und sprachlicher Räume nahelegen, so gibt es ebenso eine Reihe von Beispielen, die darauf hinweisen, dass deren Einfluss nicht überschätzt werden sollte. Beispielsweise zeigen die fünf südostasiatischen Bildungssysteme, die sowohl in TIMSS 2011 wie auch PISA 2012 die höchste durchschnittliche Mathematikleistung aufwiesen (Südkorea, Hong Kong, Japan, Taipeh, Singapur), in unserer Studie ganz unterschiedliche Leistungsprofile. Ebenso zeigen west- und südeuropäische Staaten (z. B. Belgien, Niederlande, Spanien, Italien, Zypern) heterogene Leistungsprofile, d. h. hier bieten die Daten keine Evidenz für einen Einfluss regionaler und kultureller Nähe (z. B. Europäische Union, Europäische Wirtschafts- und Währungsunion) auf nationale Curricula.