Z f Bildungsforsch (2011) 1:133–149
DOI 10.1007/s35834-011-0009-x

ORIGINALBEITRAG

zbf
ZEITSCHRIFT FÜR BILDUNGSFORSCHUNG

Faire Vergleiche? – Berücksichtigung von Kontextbedingungen des Lernens beim Vergleich von Testergebnissen aus deutschen Vergleichsarbeiten

Christiane Fiege · Franziska Reuther · Christof Nachtigall

Zusammenfassung: Eine wesentliche Säule der *Gesamtstrategie zum Bildungsmonitoring* (KMK 2006) bilden die landesweiten Vergleichsarbeiten. Diese erheben den Lern- und Leistungsstand von Schülern mittels standardisierter Tests, welche den Vergleich der Schülerleistungen zwischen verschiedenen Klassen ermöglichen. Daraus werden u. a. Aussagen über Unterrichtseffekte auf die Schülerleistung abgeleitet, die Grundlage für Unterrichtsentwicklungsmaßnahmen sein sollen. Ein Problem bei solchen Vergleichen ist, dass Klassenunterschiede nicht nur aufgrund der Unterrichtseffekte zustande kommen können, sondern auch aufgrund unterschiedlicher Ausgangsvoraussetzungen der Schüler (z. B. ihr sozioökonomischer Status). Deshalb werden bspw. einfache Mittelwertsvergleiche der Testleistungen verschiedener Klassen als unfair angesehen. Für faire Vergleiche müssen Adjustierungsverfahren verwendet werden, um diesen Unterschieden Rechnung zu tragen.

Der vorliegende Beitrag stellt die Bedeutung und Anwendung fairer Vergleiche im Kontext von deutschen Vergleichsarbeiten dar. Vor diesem Hintergrund werden die derzeit verwendeten statistischen Adjustierungsverfahren systematisiert, um sie hinsichtlich der Fairness sowie Praktikabilitätskriterien beurteilen zu können.

Angenommen: 22.03.2011 / **Online publiziert:** 14.07.2011

Dieser Artikel stellt erste Ergebnisse des Projekts „*Faire Vergleiche in der Schulleistungsforschung – Methodologische Grundlagen und Anwendung auf Vergleichsarbeiten*“ (siehe URL: http://www.fair.uni-jena.de) dar. Dieses Projekt wird vom Bundesministerium für Bildung und Forschung (BMBF) gemäß dem Rahmenprogramm zur Förderung der empirischen Bildungsforschung finanziert.

Dipl.-Psych. C. Fiege (✉) · Cand. Dipl.-Psych. F. Reuther
Lehrstuhl für Methodenlehre und Evaluationsforschung, Institut für Psychologie,
Friedrich-Schiller-Universität Jena, Projekt *Faire Vergleiche*, Jena, Deutschland
E-Mail: christiane.fiege@uni-jena.de

Cand. Dipl.-Psych. F. Reuther
E-Mail: franziska.lemke@uni-jena.de

Dr. C. Nachtigall
Lehrstuhl für Methodenlehre und Evaluationsforschung, Institut für Psychologie,
Friedrich-Schiller-Universität Jena, Projekt *kompetenztest.de*, Jena, Deutschland
E-Mail: christof.nachtigall@kompetenztest.de

Schlüsselwörter: Vergleichsarbeiten · Faire Vergleiche · Adjustierungsverfahren · Kovariaten

Fair comparisons?—Controlling for student background in German comparative performance tests

Abstract: The Standing Conference of the Ministers of Education and Cultural Affairs of the German states (KMK 2006) is currently conducting extensive monitoring of educational achievement in Germany. An important part of these efforts are the so-called "Vergleichsarbeiten" (*comparative performance tests*) that aim at assessing student achievement with standardized tests. By measuring students' achievement on one common scale, these tests allow for comparing the achievement scores of classes to assess the effects of instruction on students' outcomes. An ultimate goal of these comparisons is to identify and develop successful classroom practices. Unadjusted comparisons between classes—in the sense of naïve mean comparisons—are not fair because differences between the average achievement levels may result not only from school practice (e.g. teachers' performance) but also from pre-existing differences among students, such as socio economic status. In order to yield unbiased comparisons, adjustment procedures need to be implemented.

This article describes the significance and the implementation of fair comparisons in the context of *comparative performance testing* in Germany. Against this background, the currently implemented adjustment procedures are systematically evaluated in terms of fairness and practicability.

Keywords: Comparative performance tests · Fair comparisons · Adjustment procedures · Covariates

1 Einleitung

Nicht zuletzt aufgrund der Ergebnisse der PISA-Studie von 2000 (Baumert et al. 2001), in der Deutschland im Vergleich zu den anderen teilnehmenden Staaten in allen Fächern nur unterdurchschnittliche Leistungen erreichte, hat die Bedeutung der empirischen Bildungsforschung in den vergangenen Jahren stark zugenommen. Im Jahr 2006 beschloss die Kultusministerkonferenz die *Gesamtstrategie zum Bildungsmonitoring* (KMK 2006), welche eine systematische und wissenschaftlich fundierte Evaluation von Ergebnissen des Bildungssystems verfolgt. Diese umfasst die folgenden vier Elemente, die eng miteinander verknüpft sind, jedoch jeweils verschiedene Ebenen des Bildungssystems betreffen:

- die Teilnahme an internationalen Schulleistungsuntersuchungen,
- eine gemeinsame Bildungsberichterstattung von Bund und Ländern,
- die zentrale Überprüfung des Erreichens der Bildungsstandards im Ländervergleich sowie
- landesweite Vergleichsarbeiten.

In dem vorliegenden Beitrag fokussieren wir auf letzteren Bereich: Vergleichsarbeiten in den Bundesländern Deutschlands. Ein gemeinsames Ziel dieser Erhebungen ist die Evaluation von Unterrichtseffekten auf *Ebene einzelner Schulklassen*. Basierend auf den Testergebnissen der Schülerinnen und Schüler in den Vergleichsarbeiten sollen Maßnahmen zur Unterrichts- und Qualitätsentwicklung erarbeitet werden können, die Lehrperso-

nen einer Klasse nutzen können. Ziel dieses Beitrags ist es, eine Übersicht über die in den Bundesländern Deutschland derzeit verwendeten Auswertungsstrategien der Testergebnisse aus Vergleichsarbeiten zu erstellen.

Im Folgenden stellen wir zunächst die Bedeutung, Anwendung und Bedingungen fairer Vergleiche im Kontext von deutschen Vergleichsarbeiten dar. Vor diesem Hintergrund systematisieren wir anschließend die derzeit verwendeten Adjustierungsverfahren, um diese hinsichtlich der Fairness sowie Praktikabilitätskriterien beurteilen zu können. Dabei sei bereits an dieser Stelle angemerkt, dass sich die Auswertungsstrategien im Rahmen von Vergleichsarbeiten nicht nur zwischen den einzelnen Bundesländern unterscheiden, sondern auch im zeitlichen Verlauf verändern. Die in diesem Beitrag vorgeschlagene Systematisierung spiegelt den Stand des Jahres 2009 wider.

2 Vergleichsarbeiten in Deutschland zur Evaluation des Unterrichts

Trotz zunehmender Vereinheitlichung unterscheiden sich Vergleichsarbeiten bzgl. Testentwicklung, Testdurchführung, Datenanalyse und Ergebnisrückmeldung zwischen den einzelnen Bundesländern (vgl. Ackeren und Bellenberg 2004; Hovestadt und Kessler 2005). So existieren bspw. noch immer eine Vielzahl unterschiedlicher Bezeichnungen für die verbindlichen landesweiten Tests: Lernstandserhebungen, Kompetenztests, Orientierungsarbeiten, Diagnosearbeiten oder auch Vergleichsarbeiten (vgl. Orth 2002). Im Folgenden verwenden wir ausschließlich den letztgenannten Begriff. Vergleichsarbeiten ist gemeinsam, dass sie Schülerleistungen mittels standardisierter Testskalen erfassen. Dies ermöglicht einen Vergleich der resultierenden Testwerte. Die komparative Analyse der Testergebnisse kann dabei auf verschiedenen Ebenen stattfinden, die jeweils unterschiedliche Informationen liefern. Auf der *Schülerebene* geht es im Wesentlichen um die Lernstandsdiagnostik einzelner Schülerinnen und Schüler einer Klasse. Die Testscores sollen Informationen über ihre Stärken und Schwächen liefern, um individuelle Fördermaßnahmen seitens der Lehrperson ableiten zu können. Vergleichende Analysen auf *Klassenebene*, worauf der Fokus unseres Beitrages liegt, sollen dagegen Informationen über Unterrichtseffekte bereitstellen, die u. a. als Anstoß für Qualitätsentwicklungsmaßnahmen im Unterricht dienen sollen. Vergleichsarbeiten können somit als spezielle Evaluationsform betrachtet werden, die durch die Quantifizierung von Unterrichtseffekten Ausgangspunkt für Veränderungen im pädagogischen Handeln von Lehrkräften sein soll.

2.1 Datenanalyse bei Vergleichsarbeiten: Verortung in einen Evaluationsprozess

Damit eine Evaluation Ausgangspunkt von Veränderungen sein kann, müssen eine Reihe von Bedingungen im Prozess der Evaluation erfüllt sein. Abbildung 1 zeigt ein vereinfachtes Schema des Evaluationsprozesses im Kontext von Vergleichsarbeiten. Hier sind drei wesentliche Komponenten dargestellt: 1) die Messung, 2) die Analyse der anfallenden Daten und letztlich 3) die Rezeption der Ergebnisse seitens der Akteure im Bildungssystem, d. h. der Lehrkräfte und Schulleiter.

Die erste Komponente – *Messung* – beinhaltet die empirische Erfassung von Schülerleistungen sowie von weiteren Merkmalen des Lernumfeldes wie bspw. Schülereigen-

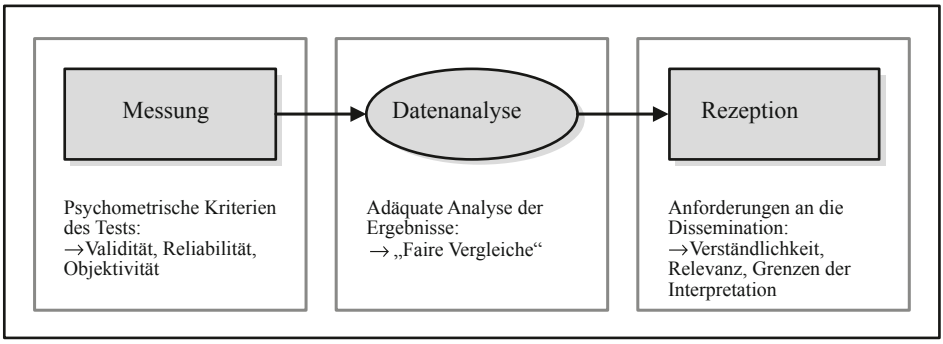

Abb. 1: Essentielle Komponenten des Evaluationsprozesses im Kontext von Vergleichsarbeiten

schaften. Hier stehen psychometrische Kriterien im Vordergrund: Die zuverlässige, valide und objektive Messung von Schülerleistungen in verschiedenen Anforderungsbereichen sowie von Kontextbedingungen des Lernens ist eine wesentliche Basis für die Vergleichbarkeit von Bildungsergebnissen. Resultat der Messung sind im Rahmen von Vergleichsarbeiten in der Regel quantitative Daten. Ein zweiter wichtiger Schritt, um die Effekte des Unterrichts zu bestimmen, ist die *Analyse der quantitativen Daten*. Die Wahl geeigneter statistischer Analyseverfahren ist ausschlaggebend für die Interpretierbarkeit der Ergebnisse als Unterrichtseffekte. Ein „fairer Vergleich“ (s. Abschn. 2.2) ist dabei die notwendige Voraussetzung zur validen Einschätzung der Wirkung von Unterricht. Die dritte Komponente bildet die *Rezeption der Ergebnisse*. Werden klassenbezogene Rückmeldungen den beteiligten Lehrkräften und Schulleitungen zur Verfügung gestellt (Dissemination), so kann nach Rolff (2002) nicht unweigerlich davon ausgegangen werden, dass hieraus unmittelbar Maßnahmen zur Unterrichtsentwicklung folgen. Erst wenn die Möglichkeiten und Grenzen der Interpretation dieser Ergebnisse verstanden werden, können Veränderungen der pädagogischen Arbeit in den Schulen erwartet werden.

Alle drei Komponenten sind gleichermaßen wichtig. Ein Großteil aktueller Forschungsbemühungen betrifft die Kompetenzmessung (z. B. Weinert 2002; Hartig und Klieme 2006; Klieme und Leutner 2006; Hartig et al. 2008; Klieme und Hartig 2008) sowie die Rezeptionsforschung(z. B. Helmke und Hosenfeld 2005; Kuper und Schneewind 2006; Maier 2008; Nachtigall et al. 2009; Müller 2010). Der Fokus dieses Beitrags liegt hingegen auf der mittleren Komponente: Die adäquate Analyse der quantitativen Daten bildet eine wesentliche Verbindung im Prozess der Evaluation von Unterrichtseffekten.

2.2 Faire Vergleiche in Vergleichsarbeiten

Wie bereits ausgeführt ist die reine Erhebung von Leistungsdaten – die Messung – mittels standardisierter Testverfahren nicht ausreichend zur Evaluation von Unterrichtseffekten. Um eine Beurteilung der aus den Vergleichsarbeiten resultierenden Ergebnisse zu ermöglichen, ist ein Vergleich der Ergebnisse mit einem Kriterium bzw. einem Standard erforderlich. Es lassen sich prinzipiell zwei Arten von Vergleichsstandards – sog. Bezugsnormen – unterscheiden, die jeweils unterschiedliche Informationen bereitstellen:

die *kriteriale* und die *soziale* Bezugsnorm (vgl. Rheinberg 2001; Watermann et al. 2003; Helmke und Hosenfeld 2004; Helmke et al. 2004). Während bei kriterialen Vergleichen ein inhaltliches Kriterium der Leistungsbeurteilung dient, erfolgt die Beurteilung einer Schülerleistung im Rahmen sozialer Vergleiche auf Basis der Leistungsverteilung aller Schülerinnen und Schüler, deren Leistung erhoben wurde.

Auch in Vergleichsarbeiten wird die soziale Bezugsnorm als Vergleichsstandard zur Beurteilung der Unterrichtseffektivität verwendet, d. h., die Testergebnisse bspw. einer Klasse werden mit den Testergebnissen anderer Klassen verglichen. Dabei liegt die Schwierigkeit darin, dass Unterschiede zwischen den Ergebnissen der Vergleichsgruppen zumeist nicht allein auf die Effektivität des Unterrichts zurückzuführen sind, sondern im Gegenteil in hohem Maße kontextabhängig sind (vgl. Nachtigall und Kröhne 2006; Nachtigall et al. 2008). Derartige Kontextvariablen des Lernens sind bspw. das Vorwissen, der sozioökonomische Status (SES) das Geschlecht, die Muttersprache eines Schülers oder die soziale Zusammensetzung der Klasse. Diese Variablen werden im Folgenden auch als Kovariaten[1] bezeichnet. Gemeinsames Merkmal von Kovariaten ist, dass sie dem Unterrichtsprozess zeitlich vorgeordnet sind. Somit sind Kovariaten von Lehrpersonen bzw. der Schule nicht beeinflussbar, haben aber ihrerseits einen Effekt auf die Schülerleistung. Ein einfacher Mittelwertsvergleich berücksichtigt solche Kontextvariablen des Lernens nicht – und verfehlt damit das Ziel, die Effektivität des Unterrichts zu quantifizieren. Es besteht daher ein allgemeiner Konsens, dass für faire Vergleiche sog. Adjustierungsverfahren verwendet werden müssen, um diesen Unterschieden Rechnung zu tragen (Watermann und Stanat 2004; Wegscheider 2004).

Adjustierungsverfahren zielen auf die Beantwortung der Frage, welches Testergebnis eine Klasse unter sonst gleichen Ausgangsbedingungen erzielt hätte, wenn eine andere Lehrkraft den Unterricht in dieser Klasse gestaltet hätte (Ceteris-paribus-Klausel; vgl. Mill 1843). Die Differenz dieses adjustierten Wertes zum tatsächlichen Testergebnis einer Klasse kann dann ursächlich auf den Effekt des Unterrichts attribuiert werden (vgl. Steyer et al. in Druck). Dies ist jedoch nur dann gerechtfertigt, wenn

- *alle* Kovariaten, die neben dem Unterricht das Testergebnis beeinflussen, in die Analyse einbezogen werden und
- das richtige statistische Modell[2] zur Analyse der Testergebnisse verwendet wird.

3 Kategorien von Adjustierungsstrategien

Derzeit gibt es im Kontext von Vergleichsarbeiten unterschiedliche statistische Adjustierungsstrategien, die das Problem der Konfundierung – d. h. der Verzerrung der geschätzten Unterrichtseffekte durch Kovariaten – zu berücksichtigen suchen. In der entsprechenden Literatur stehen bisher die inhaltlichen Ergebnisse im Fokus der Betrachtung. Die konkrete Vorgehensweise zur Berechnung der Vergleichswerte und deren methodische Fundierung werden zumeist nur unzureichend dokumentiert. Aufgrund aktueller Rechercheergebnisse (Quellenanalyse vorhandener Literatur bzw. von Internetquellen; Stand: Dezember 2009) bundeslandspezifischer Vergleichsarbeiten wurde in Anlehnung an Nachtigall et al. (2008) und Fiege (2007) eine Systematik von Adjustierungsstrategien erstellt, der die

Tab. 1: Kategorien von Adjustierungsstrategien in deutschen Vergleichsarbeiten

Strategie		Anmerkung	Beschreibung des Referenzwertes	Beispiele
I	Vergleich mit Landesmittelwert	Unadjustierte Vergleiche	Landesmittelwert	VERA 3 in Sachsen-Anhalt
II	Vergleich mit subgruppenspezifischem Mittelwert	Marginale Adjustierung: Subklassifikation	Mittelwert innerhalb einer Subpopulation (bspw. Schulart, Geschlecht)	VERA 8 in Brandenburg
III	Vergleich mit ähnlichen (existierenden) Klassen	IIIa Standorttypen	Auswahl von Schulen des gleichen Standorttyps	Lernstand 8 in Nordrhein-Westfalen
		IIIb Belastungsindex	Auswahl von vier Schulen mit ähnlichstem Belastungsindex	Lernstand 8 in Hamburg
		IIIc Kontextgruppen	Auswahl von Schulen der gleichen Kontextgruppe	VERA 3 in Rheinland-Pfalz
IV	Vergleich mit Erwartungswert	Outcome-Modellierung	Regressionsanalytisch vorhergesagter Wert unter Berücksichtigung verschiedener Kovariaten (Geschlecht, SES, Muttersprache etc.)	Kompetenztest in Thüringen

verschiedenen Vergleichsarbeiten der einzelnen Bundesländer zugeordnet werden können. Gemeinsames Merkmal dieser Strategien ist, dass sich die Adjustierung jeweils auf die Berechnung des Referenzwertes bezieht. Unabhängig von der Art der Adjustierung können dabei Mittelwerte von Testleistungen, Lösungshäufigkeiten, Kompetenzniveauverteilungen oder andere Verteilungskennwerte als Referenzwert dienen. Tabelle 1 stellt die vier Kategorien von Referenzwerten dar, die sich im Rahmen der Adjustierung der Testergebnisse aus den bundeslandspezifischen Vergleichsarbeiten unterscheiden lassen.

Die einzelnen Kategorien von Adjustierungsstrategien können hinsichtlich verschiedener Kriterien charakterisiert werden, die in einem engen Zusammenhang miteinander stehen: Fairness, Testökonomie und Modellkomplexität.

Fairness. Erst durch die Berücksichtigung von Kovariaten, d. h. Kontextmerkmalen des Lernens, auf die die Lehrperson einer Klasse keinen Einfluss hat, sind faire Vergleiche möglich. Eine notwendige Bedingung fairer Vergleiche – im Sinne der obigen Definition (s. Abschn. 2.2) – besteht darin, dass *alle* Kovariaten in die Analyse einbezogen werden müssen. Dies ist im schulischen Kontext häufig nicht realisierbar, denn bspw. aus testökonomischen Gründen können nicht alle Einflussfaktoren des schulischen Lernens erhoben werden. Daher versucht man, zumindest einige wichtige Einflussvariablen (wie bspw. Schulart, Geschlecht, sozioökonomischer Status etc.) zu berücksichtigen. In diesem Sinne stellen Adjustierungsverfahren in der Praxis empirischer Bildungsforschung in der Regel eine Annäherung an faire Vergleiche dar. Häufig wird daher auch von „faireren" Vergleichen gesprochen (vgl. Nachtigall et al. 2009, S. 9). Hier unterscheiden sich die Adjustierungsstrategien hinsichtlich der Art (Welche Kovariaten werden berücksichtigt?)

und Anzahl (Wie viele Kovariaten fließen in die Analyse ein?) der berücksichtigen Kovariaten. In diesem Zusammenhang spielt immer auch die messtheoretische Qualität bzgl. der Erfassung der Kovariaten eine wichtige Rolle: Erst eine reliable, valide und objektive Messung der Kovariaten ermöglicht die adäquate Berücksichtigung dieser im Rahmen von Adjustierungsverfahren.

Testökonomie. Ein weiteres Kriterium – die *Testökonomie* (vgl. Moosbrugger und Kelava 2007) – bezieht sich auf die Erfassung der in der Analyse berücksichtigten Kovariaten. Während in einigen Vergleichsarbeiten Daten der amtlichen Statistik genutzt werden, verwendet man in anderen Bundesländern zusätzliche Fragebögen für Schüler und Lehrer, um Informationen hinsichtlich relevanter Kovariaten zu erheben. Letztere Vorgehensweise setzt jedoch die Motivation und auch zeitliche Ressourcen der Schüler und Lehrer voraus, zusätzlich zu dem ohnehin schon aufwändigen Testverfahren. Testökonomie bezieht sich also auf die Frage, ob die relevanten Kontextvariablen sparsam, ohne große zusätzliche Kosten (wie Zeit, Geld oder andere Ressourcen) erfasst werden können.

Modellkomplexität. Auch hinsichtlich des methodischen Vorgehens bei der Datenanalyse lässt sich ein Sparsamkeitskriterium differenzieren. Dabei geht es um die *Komplexität des Modells*, also die Frage, ob zur Berechnung des adjustierten Vergleichswertes komplexe statistische Modelle angewendet werden. Die Modelle sollten dabei so komplex wie nötig sein, d. h., sie sollen die tatsächlich bestehenden Zusammenhänge der Variablen adäquat abbilden können. Dies ist die zweite notwendige Bedingung für faire Vergleiche (s. Abschn. 2.2). Die Modelle sollten aber auch so einfach wie möglich sein, da in komplexen statistischen Modellen mehr Parameter geschätzt werden müssen. Dadurch steigen die Anforderungen an die zugrundeliegenden Daten, denn um diese Parameter schätzen zu können, sind z. T. zusätzliche Annahmen bzw. größere Stichprobenumfänge nötig. Die Beobachtungsanzahl ist im schulischen Kontext jedoch natürlich begrenzt durch die Anzahl der Schülerinnen und Schüler einer Klasse, einer Schule bzw. eines Bundeslandes. Des Weiteren können komplexe statistische Modelle von einem methodisch wenig geschulten Publikum nicht ohne Weiteres verstanden werden. Für die Rezeption und Verwertung der Ergebnisse im Rahmen von Unterrichtsentwicklungsmaßnahmen kann dies nachteilig sein, da die fehlende Transparenz bzgl. der Datenanalysen potentiell zu Unverständnis und verminderter Akzeptanz führt (vgl. Braun et al. 2010).

Die drei Kriterien – Fairness, Testökonomie und Komplexität des Modells – stellen die wesentlichen Dimensionen dar, die im Folgenden zur Charakterisierung der Adjustierungskategorien genutzt werden.

3.1 Strategie I: Vergleich mit dem Landesmittelwert

Diese erste Strategie der Datenanalyse ist dadurch gekennzeichnet, dass die jeweiligen Testwerte mit dem Landesmittel, d. h. dem Mittelwert der Testwerte aller Schülerinnen und Schüler einer Klassenstufe innerhalb eines Bundeslandes, verglichen werden. Dieses Vorgehen wird bspw. in Sachsen-Anhalt im Rahmen von VERA 3 praktiziert (siehe URL: http://www.bildung-lsa.de/home.html). Hier werden die Landesergebnisse – also die durchschnittliche Testleistung aller Grundschüler in Sachsen-Anhalt – auf dem Bildungsserver des Landes Sachsen-Anhalt veröffentlicht. Diese können von den Lehrkräften der

einzelnen Klassen als Referenz bei der schulinternen Leistungsbeurteilung genutzt werden. Eine Adjustierung des Referenzwertes basierend auf der Berücksichtigung wichtiger Hintergrundvariablen des Lernens findet bei dieser Strategie der Datenauswertung nicht statt. Somit sind auch keine fairen Vergleiche möglich. Eine ursächliche Attribution gefundener Unterschiede als Unterrichtseffekte ist somit nicht gerechtfertigt. Differenzen zum Landesdurchschnitt können lediglich im Rahmen einer sozialen Bezugsnorm interpretiert werden, ohne dabei Informationen über deren Ursachen zu liefern. Zwar wird in Sachsen-Anhalt hinsichtlich der pädagogischen Nutzung der Ergebnisse darauf hingewiesen, dass Unterschiede in den Testwerten auf unterschiedliche Ursachen zurückzuführen sind. So sollen sich die individuellen Lehrkräfte bei der Auswertung der Ergebnisse als Leitfrage u. a. damit befassen, wo mögliche Ursachen für gefundene Differenzen liegen. Dennoch besteht die Gefahr, dass derartige Unterschiedswerte der Unterrichtsqualität zugeschrieben werden. Besonders in Klassen, die über dem jeweiligen Landesdurchschnitt liegen, ist dann eine differenzierte Beurteilung der Unterrichtsqualität unwahrscheinlich. Hinsichtlich der Testökonomie ist diese Vorgehensweise ebenfalls nachteilig zu bewerten. Zwar fallen, neben dem zeitlichen Aufwand für die Testung der Vergleichsarbeiten, keine weiteren zeitlichen oder sonstigen Kosten an. Jedoch stehen auch keine diagnostischen Informationen zur Zusammensetzung der Schüler hinsichtlich wichtiger Kovariaten zur Verfügung.

3.2 Strategie II: Vergleich mit einem subgruppenspezifischen Mittelwert

Die Rückmeldung eines unadjustierten Vergleichswertes ist also offensichtlich nicht ausreichend, insbesondere dann nicht, wenn unterschiedliche Schularten in den Vergleich einfließen. So wird bspw. VERA 8 in Brandenburg an Gymnasien, Gesamtschulen, Oberschulen und Förderschulen durchgeführt (Emmrich et al. 2010). Hier findet eine sog. *marginale Adjustierung* statt, d. h., der durchschnittliche Testwert innerhalb der jeweiligen Subpopulation wird als Vergleichswert betrachtet. Der mittlere Testwert einer Gymnasialklasse wird also nicht dem mittleren Testwert aller Schüler Brandenburgs, sondern dem Mittelwert aller Gymnasiasten der gleichen Klassenstufe gegenübergestellt (Emmrich 2010). Diese marginalen Adjustierungen können sich, neben der Schulform, auch auf andere Variablen beziehen. So meldet das ISQ (Institut für Schulqualität der Länder Berlin und Brandenburg e. V.; siehe URL: http://www.isq-bb.de/) in Brandenburg zusätzlich geschlechtsspezifische Klassen- und Landesmittelwerte innerhalb einer Schulform zurück, die miteinander verglichen werden können.

Obgleich auch diese zweite Strategie keine fairen Vergleiche – im Sinne von Unterrichtseffekten – liefert, stellt sie hinsichtlich der drei formulierten Kriterien eine Weiterentwicklung gegenüber Strategie I dar: Die Ergebnisse sind *fairer*, da relevante Kovariaten wie Schulform und Geschlecht in der Analyse berücksichtigt werden. Diese Kovariaten können ökonomisch erfasst werden, da sie in der Regel als Schülerstammdaten bei der Durchführung der Vergleichsarbeiten per se erhoben werden. Des Weiteren sind die Ergebnisse leicht verständlich und gut kommunizierbar, da lediglich subgruppenspezifische Mittelwerte berechnet werden und kein komplexes statistisches Modell Grundlage der Analyse ist.

3.3 Strategie III: Vergleich mit ähnlichen (existierenden) Klassen

Charakterisierende Eigenschaft der dritten Strategie ist die Konstruktion einer Kovariaten, welche die soziale Belastung einzelner Klassen bzw. Schulen quantifiziert (aktuell hierzu vgl. Bonsen et al. 2010). Dabei werden Informationen aus jeweils verschiedenen Indikatoren für den sozialen Hintergrund der Schülerinnen und Schüler einer Klasse zu einem Kennwert aggregiert. Die dabei resultierenden Vergleichswerte sind im Vergleich zu den ersten beiden Strategien als *fairer* zu betrachten, da neben der Schulform auch der soziale Hintergrund eines Schülers Berücksichtigung findet. Dieser stellt eine der wichtigsten Kovariaten im Schulleistungskontext dar (vgl. Baumert und Schümer 2001). Aus der methodischen Perspektive lassen sich drei Substrategien differenzieren, die sich in der Komplexität des methodischen Vorgehens, aber auch hinsichtlich der Testökonomie unterscheiden.

Strategie IIIa. In Nordrhein-Westfalen wird bei der Lernstandserhebung in Klasse 8 eine sog. *Standorttypisierung* zur Erstellung fairer Vergleiche genutzt. Die Standorttypen charakterisieren die soziodemographische Zusammensetzung der Schülerschaft innerhalb der verschiedenen Schulformen Hauptschule, Gesamtschule, Realschule und Gymnasium sowie regionale Merkmale des Schulstandorts. Die Definition der Standorttypen erfolgte schulformbezogen: Bei den Haupt- und Gesamtschulen wurden drei, bei den Realschulen und Gymnasien zwei Standorttypen gebildet. „Die Beschreibungen der Typen [bei der Lernstandserhebung 2004] konzentrierten sich auf Merkmale, die sich in der Pilotstudie zu den Lernstandserhebungen und auch in der bundeslandspezifischen NRW-Auswertung der PISA-Daten als solche erwiesen haben, die einen hohen Zusammenhang mit den erzielten Fachleistungen der Schülerinnen und Schüler in den jeweiligen Schulformen aufweisen“ (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen 2005, S. 3). Derzeit wird dieses Standorttypenkonzept in Nordrhein-Westfalen in Kooperation mit dem Landesamt für Datenverarbeitung und Statistik weiterentwickelt (K. Isaac, persönl. Mitteilung, 27.08.2009). Im Rahmen der Durchführung der jährlichen Lernstandserhebung ist die Schulleitung aufgefordert, die Zuordnung ihrer Schule – und damit auch der getesteten Klassen – zu einem der Standorttypen vorzunehmen. Zu diesem Zweck wird durch die Schulleitung eine Checkliste bearbeitet, die für jeden der Standorttypen verschiedene Kriterien auflistet. Abschließend soll der Standorttyp ausgewählt werden, bei dem die meisten dieser Kriterien zutreffen. Die Erfassung bzw. Zuordnung des Standorttyps der einzelnen Schulen ist hier also sehr ökonomisch gestaltet. Allerdings ist die messtheoretische Qualität in diesem Zusammenhang nicht ganz unproblematisch, da die Zuordnung durch die Schulen selbst erfolgt. Hierbei ist bspw. nicht auszuschließen, dass sich Schulen einem Standorttyp mit schlechteren Ausgangvoraussetzungen zuordnen, um bei dem Vergleich selbst besser abzuschneiden. Die Rückmeldungen enthalten dann u. a. die Kompetenzniveauverteilung von allen Schulen des gleichen Standorttyps als Referenz.

Strategie IIIb. Auch im Rahmen von Strategie IIIb wird der in einer Klasse erreichte durchschnittliche Testwert dem mittleren Ergebnis mehrerer ähnlicher – also vergleichbarer – Klassen gegenübergestellt. Die Referenzgruppe wird hier jedoch hinsichtlich eines metrischen Indikators für den sozioökonomischen Status, dem sog. *Belastungsindex* einer Schule, ausgewählt. Ein typischer Repräsentant dieses Vorgehens war bis

2009 die Lernstandserhebung 8 in Hamburg (siehe URL: http://www.lernstand.hamburg.de/)[3]. Hier wurden die vier Schulen der gleichen Schulform, die den der eigenen Klasse ähnlichsten Kennwert bzgl. des Belastungsindex aufwiesen, als Referenzgruppe bei der Darstellung der Leistungsverteilung vergleichbarer Klassen berichtet. Die Konstruktion des Belastungsindex für den Sekundarschulbereich erfolgte dabei im Rahmen der KESS 7-Studie auf der Basis eines komplexen statistischen Verfahrens (vgl. Bos et al. 2007), wobei verschiedene Indikatoren des sozialen Hintergrundes mittels eines Rasch-Modells (Rasch 1960) analysiert wurden. Die Informationen zum sozialen Hintergrund wurden im Rahmen von Schüler- und Elternbefragungen erhoben.

Dieses Vorgehen ist sehr testökonomisch, da die Zuordnung von Klassen zu Werten des Sozialindikators nicht mit jeder Erhebung erneut vorgenommen werden muss[4]. Diese Strategie ist auch insofern vorteilhaft, als die jeweiligen Vergleichswerte einfach zu berechnen und leicht zu interpretieren sind. Dem steht jedoch der Nachteil gegenüber, dass bei Betrachtung von nur vier Vergleichsschulen zufallsbedingte Schwankungen zu erwarten sind: Einzelne Schulen mit größerem Anteil sozial benachteiligter Schüler können u. U. im Mittel auch höhere Vergleichswerte erzielen als Klassen mit weniger benachteiligten Schülern, ohne dass dieser Unterschied auf Unterrichtseffekte zurückzuführen ist. Ein weiterer Kritikpunkt betrifft die Messung der Indikatoren des sozialen Hintergrundes. Hier bilden, neben den Schülerangaben, auch Daten aus Elternbefragungen die Grundlage zur Berechnung des Belastungsindex. An dieser Stelle kann das Problem der Unit-Nonresponse (vgl. Schafer und Graham 2002) auftreten, wenn Eltern bspw. aufgrund mangelnder Motivation oder anderer Ursachen die Teilnahme an der Befragung ablehnen. Dies kann potentiell zu systematischen Verzerrungen bei der Berechnung der Indizes führen, wenn vor allem Eltern sozial benachteiligter Familien nicht an der der Befragung teilnehmen.

Strategie IIIc. Eine methodisch komplexe Analysestrategie stellt auch das Kontextuierungsverfahren dar, welches im Rahmen des Projekts „Vergleichsarbeiten in der Grundschule“ (Projekt VERA[5], siehe URL: http://www.projekt-vera.de) entwickelt wurde. Hier werden für die Erstellung fairer Vergleiche sog. *Kontextgruppen* gebildet (vgl. Isaac und Hosenfeld 2008). Diese werden in einem ersten Analyseschritt anhand einer repräsentativen Zufallsstichprobe für jedes Bundesland (sog. „Zentralstichprobe“) im Rahmen eines regressionsanalytischen Mehrebenen-Ansatzes (Hierarchisch Lineare Modellierung, HLM) ermittelt. Dabei wird der Gesamtleistungswert, d. h. der Mittelwert der Schülerleistungen in den beiden getesteten Fächern Deutsch und Mathematik, durch mehrere Prädiktoren auf Schüler- und Klassenebene (bspw. Geschlecht, Anteil von Jungen, Anteil von Kinder aus Familien mit Arbeitslosigkeit etc.) vorhergesagt. Nur die Prädiktoren, welche sich als statistisch signifikant erweisen, fließen in die weiteren Berechnungen ein. Die für die einzelnen Klassen der Zentralstichprobe aufgrund dieser Prädiktoren vorhergesagten Leistungswerte sind die sog. *Kontextwerte*. Die Verteilung der Kontextwerte wird anschließend in drei Gruppen eingeteilt: Kontextgruppe I bzw. III enthält jeweils 25 % aller Klassen mit den niedrigsten bzw. höchsten Kontextwerten und die mittlere Kontextgruppe II schließt 50 % der Klassen ein. In einem zweiten Schritt erfolgt Zuordnung aller Klassen zu den Kontextgruppen aufgrund von Lehrerangaben im Rahmen der Durchführung der jährlichen Vergleichsarbeiten. Dazu werden die Lehrkräfte gebeten, neben der Eingabe der Testleistungsdaten im geschützten online-Bereich des VERA-Pro-

jekts, Einschätzungen zum sozialen Hintergrund ihrer Klasse anzugeben. Diese Daten dienen, zusammen mit den Schülerstammdaten, zur Bestimmung des Kontextwertes mittels der an der Zentralstichprobe bestimmten Regressionsgewichte und somit der Zuordnung der einzelnen Klassen zu einer der drei Kontextgruppen. Als klassenspezifische Referenz wird dann für jeden Inhaltsbereich (wie bspw. Deutsch: Lesen) die Kompetenzniveauverteilung aller Klassen aus der eigenen Kontextgruppe zurückgemeldet.

Ähnlich der Strategie IIIb basiert auch die Konstruktion der Kontextgruppen auf einem methodisch komplexeren Analyseverfahren. Im Gegensatz zum Vorgehen in Hamburg ist dieses Verfahren jedoch weniger testökonomisch, da zur Ermittlung der Kontextgruppe einer Klasse jedes Jahr zusätzliche Angaben seitens der Lehrperson notwendig sind: Werden die Zusatzfragebögen aufgrund mangelnder Zeit oder Motivation nicht ausgefüllt, können in den klassenspezifischen Rückmeldungen keine fairen Vergleiche angegeben werden. Zudem ergeben sich ähnliche messtheoretische Probleme wie in Strategie IIIa: Auch hier kann es zu einem strategischen Antwortverhalten kommen, wenn Lehrkräfte den sozialen Hintergrund ihrer Klasse schlechter einschätzen, um in einem Vergleich besser abzuschneiden.

3.4 Strategie IV: Vergleich mit einem Erwartungswert

Bei der vierten Strategie werden Vergleiche nicht bzgl. tatsächlich existierender Klassen durchgeführt, sondern bzgl. eines Erwartungswertes. Dieser Erwartungswert stellt den für eine Klasse mit ähnlichen Kontextbedingungen hinsichtlich relevanter Schülermerkmale zu erwartenden Leistungswert bzw. durchschnittlichen Testwert dar. In diese Strategie lässt sich das Adjustierungsverfahren, welches im Rahmen der Thüringer Kompetenztests der Klassenstufen 3, 6 und 8 (Projekt „kompetenztest.de“, siehe URL: http://www.kompetenztest.de/) praktiziert wird, einordnen. Hier wird zunächst für jeden Schüler einer Klasse ein Erwartungswert berechnet. Diese Erwartungswerte werden über die Zellenmittelwerte einer multifaktoriellen Varianzanalyse (ANOVA-Zellenmittelwertemodell) geschätzt (vgl. Nachtigall et al. 2008). Abhängige Variable ist dabei die Testleistung der Schülerinnen und Schüler. Die verschiedenen Faktoren sind folgende diskrete bzw. diskretisierte Kovariaten: Schulart, Geschlecht, Diagnose besonderer Lernschwierigkeiten bzw. sonderpädagogischer Förderbedarf, Wiederholer einer Klassenstufe, Muttersprache, Anzahl der Bücher im Elternhaus (als Indikator des SES) sowie der Vortestwert[6], d. h. das Testergebnis aus dem Kompetenztest einer früheren Klassenstufe. Der Erwartungswert eines individuellen Schülers ist somit der Mittelwert aller Schülerinnen und Schüler mit der gleichen Kovariatenkonstellation, d. h. den gleichen Ausprägungen auf den berücksichtigten Kovariaten[7]. Der adjustierte Vergleichswert einer Klasse, der sog. *korrigierte Landesmittelwert*, ist dann der Mittelwert der geschätzten schülerspezifischen Erwartungswerte der jeweils betrachteten Klasse.

Hinsichtlich des Fairness-Kriteriums besteht der Zugewinn gegenüber den vorangegangenen Vorgehensweisen (Strategien I bis III) darin, dass im Rahmen der Thüringer Kompetenztests eine weitere relevante Kovariate berücksichtigt wird: Neben den Stammdaten – wie Geschlecht und Schulart – und dem sozioökonomischen Hintergrund wird zusätzlich auch der Vortestwert der Schüler aus Kompetenztests früherer Klassenstufen in die Analyse einbezogen[8]. Dieser Vortestwert kann als Indikator für das bereichsspe-

zifische Schülervorwissen betrachtet werden, welches eine zentrale Determinante der Schülerleistungen darstellt (vgl. Hedges und Hedberg 2007; Nachtigall et al. 2008; Schrader und Helmke 2008).

Die Berücksichtigung des Vorwissens setzt das Vorliegen längsschnittlicher Daten voraus. Damit lässt sich die im Projekt „kompetenztest.de“ verwendete Adjustierungsstrategie in die Tradition der *Value-Added-Modellierung* einordnen, deren gemeinsames Merkmal die Modellierung längsschnittlicher Schülerleistungsdaten ist. Der Begriff Value-Added-Modelle (VAM) bezieht sich auf eine ganze Familie verschiedener statistischer Modelle, deren Ziel die Quantifizierung von Schul- bzw. Unterrichtseffekten ist (vgl. Braun und Wainer 2007, S. 867). Hierzu zählen auch Kovariaten-Adjustierungsmodelle, welche die aktuelle Testleistung als Funktion der Vortestwerte sowie weiterer Kovariaten spezifizieren (vgl. McCaffrey et al. 2003, S. 55). Ein solches Modell findet auch im Rahmen der hier beschriebenen Thüringer Kompetenztests Anwendung. An dieser Stelle sei noch einmal betont, dass die Verwendung des Vortestwertes kein genuines Merkmal der Strategie IV (Vergleich mit einem Erwartungswert) darstellt, sondern eine Besonderheit des Vorgehens im Projekt „kompetenztest.de“ ist. Bei der Interpretation der auf diese Weise berechneten Effektivitätsmaße ist allerdings zu berücksichtigen, dass der Erhebungszeitpunkt des Vortests eine zentrale Rolle spielt. Wird in Klassenstufe 8 bspw. der Vortest aus Klassenstufe 6 berücksichtigt, werden auch die bis zur sechsten Klasse erreichten Unterrichtseffekte auspartialisiert. In diesem Fall würde man also den Effekt des Unterrichts in den Klassenstufen 7 bis 8 betrachten. Bezogen auf die Testökonomie unterscheidet sich das Vorgehen bei den Thüringer Kompetenztests nicht von Strategie IIIc, denn auch hier müssen die Lehrpersonen Zusatzangaben machen, um die Berechnung des adjustierten Landesmittelwertes zu ermöglichen. Dies birgt vergleichbare messtheoretische Probleme, die bereits in den Strategien IIIa bis IIIc beschrieben wurden und u. U. zu systematischen Verzerrungen führen können. Das der Strategie IV zugrundeliegende statistische Modell – ein ANOVA-Zellenmittelwertemodell – ist ähnlich komplex wie die in den Strategien IIIb und IIIc verwendeten Modelle.

4 Diskussion

4.1 Grenzen und Möglichkeiten von fairen Vergleichen in Vergleichsarbeiten

Landesweite Vergleichsarbeiten gehören mittlerweile zum Standardrepertoire der empirischen Bildungsforschung in den Bundesländern Deutschlands. Sie sind seit dem Jahr 2006 auch Teil der KMK-Gesamtstrategie zum Bildungsmonitoring des deutschen Bildungssystems. Die Ergebnisse dieser standardisierten Testverfahren sollen u. a. Aussagen über Unterrichtseffekte ermöglichen und Ausgangspunkt für Unterrichtsentwicklung sein können. Dazu werden die Testergebnisse einer Klasse mit den Ergebnissen anderer Klassen verglichen (soziale Bezugsnorm). Hier müssen jedoch Kontextfaktoren, die ebenfalls die Schülerleistung beeinflussen (Kovariaten), in der Analyse der Testergebnisse berücksichtigt werden, um faire Vergleiche zu ermöglichen.

Obwohl bezüglich der Ziele ein einheitlicher Rahmen vorliegt, zeigt sich bei näherer Betrachtung der Vergleichsarbeiten über die einzelnen Bundesländer hinweg ein sehr

heterogenes Bild. Unterschiede bestehen nicht nur in Bezug auf die Bezeichnung der Vergleichsarbeiten, sondern auch im Hinblick auf die Testentwicklung, Testdurchführung, Testauswertung sowie die Rückmeldung der Ergebnisse.

Der Fokus dieses Beitrags lag auf der Analyse der Testergebnisse, welche als ein zentrales Bindeglied zwischen der Messung (Erfassung von Schülerleistungen und Kontextfaktoren des Lernens) und der Ergebnisrezeption fungiert. Auch für die Analysestrategien im Rahmen der Vergleichsarbeiten zeigt sich über die Bundesländer ein uneinheitliches Bild, welches jährlichen Änderungen unterliegt. Partiell werden noch unadjustierte Vergleichswerte zurückgemeldet. Werden Adjustierungen durchgeführt, so beziehen sich diese auf die Berechnung des Vergleichswertes. Die Differenz des beobachteten Klassenmittelwertes vom jeweils berechneten adjustierten Vergleichswert soll dann als Maß der Effektivität des Unterrichts interpretiert werden können. Aus methodischer Sicht ist dabei der Fairness-Aspekt zentral: Nur wenn *alle* Kovariaten in der Analyse adäquat, d. h. mit Hilfe des richtigen statistischen Modells, berücksichtigt werden, können die berechneten Differenzwerte als ursächliche Effekte des Unterrichts interpretiert werden. Dies setzt zudem die messtheoretische Qualität (Reliabilität, Validität und Objektivität) bei der Erfassung der Kovariaten voraus. In der Praxis empirischer Bildungsforschung spielen jedoch stets auch Praktikabilitätsaspekte wie Testökonomie oder Modellkomplexität eine nicht zu vernachlässigende Rolle. So können bspw. aus testökonomischen Gründen nicht alle Kovariaten erfasst werden. Auch das verwendete statistische Modell muss einem Sparsamkeitskriterium genügen, nicht zuletzt um die Transparenz und Kommunizierbarkeit der berechneten Effektgrößen zu ermöglichen. Aus diesen Gründen können die berechneten Effektmaße stets nur eine Annährung an Unterrichtseffekte – und in diesem Sinne lediglich *fairere* Vergleiche darstellen. Daher sollten die berechneten Maße nicht als ursächliche – also kausale – Effekte des Unterrichts, sondern als deskriptive Maße interpretiert werden (vgl. Briggs 2008). Trotz der erwähnten Einschränkungen bergen solche faireren (d. h. adjustierten) Vergleichswerte ein großes Potential, als Informationsbasis für die beteiligten Lehrkräfte zu dienen. Briggs (2008, S. 12) begründet diese Auffassung wie folgt: „Given a quasi-experimental design, a VAM [value-added model] may be the closest we can come to an approximation of this ideal [i.e., estimating causal effects of instruction]“. Die so berechneten Vergleichswerte können Ausgangspunkt für Diskussionen sein sowie Impulse für Unterrichtsentwicklungsmaßnahmen geben, wenn die Möglichkeiten, aber auch die Grenzen der Interpretation transparent dargestellt werden (vgl. Maier 2008).

4.2 Mindestanforderungen für Low-Stakes Assessment

Dennoch sollte ein gemeinsamer Standard für die Bundesländer Deutschlands, der die oben genannten Kriterien fairer Vergleiche berücksichtigt, als Grundlage bei der Analyse von Daten aus Vergleichsarbeiten (und der Ergebnisdissemination) dienen. Die Heterogenität des Vorgehens bei der Datenanalyse zwischen den Bundesländern spiegelt diesbezüglich einen Missstand wider. Es erscheint zunächst willkürlich, welche Analysestrategien angewendet werden. Insbesondere die Auswahl der Kovariaten hat jedoch einen empfindlichen Einfluss auf die Ergebnisse. So können adjustierte Vergleiche für einzelne Klassen vollkommen unterschiedlich ausfallen, je nachdem, welche Kovariaten in der

Adjustierung verwendet werden (Fiege et al. 2010). Auch die Art des dabei geschätzten Effekts (also Unterrichtseffekt, Lehrereffekt oder Schuleffekt) ist abhängig von der Wahl der Kovariaten (vgl. Raudenbush und Willms 1995; Meyer 1997). Rankings oder Sanktionierungsmaßnahmen (wie die Kürzung von Lehrergehältern im Rahmen von High-Stakes Assessment Systemen bspw. in den USA) können vor diesem Hintergrund nicht gerechtfertigt werden. Aber auch Low-Stakes Assessments, die keine derart drastischen Konsequenzen für die Lehrpersonen einer Klasse haben, sollten einen gemeinsamen Standard als Grundlage haben. Ein verantwortungsvoller Umgang mit den Testergebnissen setzt voraus, dass die Analyse der Testdaten methodischen Mindestanforderungen genügt und transparent dargestellt wird. Dies sollte unserem Erachten nach bei der Umsetzung der KMK-Gesamtstrategie zum Bildungsmonitoring des deutschen Bildungssystems, insbesondere im Rahmen der hier dargestellten Vergleichsarbeiten, zukünftig noch stärker berücksichtigt werden. Die in diesem Beitrag dargestellten Kriterien der Fairness und Praktikabilität können dabei als hilfreiche Grundlage dienen. Weitere Schritte zur Entwicklung methodischer Richtlinien sollten jedoch auch empirische Untersuchungen der Adjustierungsverfahren, sowie die weitere Integration der Forschungsergebnisse zu den in anderen Ländern angewendeten Value-Added-Modellen sein.

Anmerkungen

1 Der Begriff Kovariaten bezieht sich im Folgenden sowohl auf individuelle Merkmale von Schülern, als auch auf Kontextvariablen.

2 Im Rahmen von Adjustierungsverfahren werden Zusammenhänge zwischen Variablen – hier der Testleistung der Schüler und den Kovariaten – mittels mathematischer Modelle dargestellt. Voraussetzung dabei ist, dass das gewählte statistische Modell die tatsächlichen Zusammenhänge zwischen diesen Variablen abbildet. Dabei gibt es eine Vielzahl mathematischer Modelle, deren detaillierte Darstellung jedoch weit über den Rahmen dieses Artikels hinaus geht. Der interessierte Leser sei auf weiterführende Literatur wie bspw. Bortz (2005) verwiesen.

3 Ab dem Jahr 2010 werden im Rahmen der Rückmeldungen zu Lernstand 8 in Hamburg die schulartspezifischen Mittelwerte (also bspw. die mittlere Testleistung aller Haupt- und Realschulen) als Referenzwerte zurückgemeldet. Außerdem kann sich eine Schule hinsichtlich der schulartspezifischen Kompetenzniveauverteilungen vergleichen (F. Thonke, persönl. Mitteilung, 08.06.2010). Dieses Vorgehen entspricht nun also dem Vorgehen gemäß Strategie II.

4 Dabei wird die Annahme gemacht, dass der Einzugsbereich einer Schule und der soziale Hintergrund der Schüler über mehrere Jahre relativ stabil bleibt (vgl. Freie und Hansestadt Hamburg 2009).

5 Das Projekt VERA in Landau wertet derzeit die Testergebnisse der Vergleichsarbeiten in Klassenstufe 3 für insgesamt acht Bundesländer aus (Baden-Württemberg, Bremen, Mecklenburg-Vorpommern, Niedersachen, Nordrhein-Westfalen, Rheinland-Pfalz, Saarland, Schleswig-Holstein). Für Klassenstufe 8 werden die Testergebnisse aus vier Bundesländern (Bremen, Niedersachen, Rheinland-Pfalz, Saarland) ausgewertet.

6 Das Projekt „kompetenztest.de“ wertet auch die Testergebnisse der Vergleichsarbeiten in Hessen, Mecklenburg-Vorpommern und Sachsen mit diesem Verfahren aus. Allerdings liegen in diesen Ländern keine Längsschnittdaten vor, so dass nur in Thüringen der Vortestwert als Kovariate berücksichtigt werden kann.

7 Das Vorgehen in Strategie IV ist bis zu diesem Analyseschritt identisch zu Strategie II, wobei in letzterer zumeist nur die Kovariaten Schulart und Geschlecht berücksichtigt werden. Zudem werden in Strategie II die auf diese Weise adjustierten Werte nicht auf Klassenebene aggregiert, sondern pro Kovariatenkonstellation zurückgemeldet. So wird im Rahmen von Strategie II bspw. dem Testleistungsmittelwert der Mädchen einer Gymnasialklasse der Mittelwert aller anderen weiblichen Gymnasiasten vergleichend gegenübergestellt.

8 Der Vortestwert eines Schülers oder einer Schülerin kann nur dann in die Analyse einbezogen werden, sofern diese Information zur Verfügung steht. Für Individuen, die bspw. aufgrund eines Umzugs erst ab Klassenstufe 8 Teil des Thüringer Schulsystems sind, liegen keine Daten aus früheren Kompetenztests vor.

Literatur

Ackeren, I. van, & Bellenberg, G. (2004). Parallelarbeiten, Vergleichsarbeiten und Zentrale Abschlussprüfungen – Bestandsaufnahme und Perspektiven. In H. G. Holtappels, K. Klemm, H. Pfeiffer, H.-G. Rolff, & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (S. 125–160). Weinheim: Juventa.

Baumert, J., & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In J. Baumert et al. (Hrsg.), *PISA 2000– Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 323–410). Opladen: Leske + Budrich.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (Hrsg.). (2001). *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Bonsen, M., Bos, W., Gröhlich, C., Harney, B., Imhäuser, K., Makles, A., Schräpler, J.-P., Terpoorten, T., Weishaupt, H., & Wendt, H. (2010). *Zur Konstruktion von Sozialindizes – Ein Beitrag zur Analyse sozialräumlicher Benachteiligung von Schulen als Voraussetzung für qualitative Schulentwicklung*. Bildungsforschung Band 31, Herausgegeben vom Bundesministerium für Bildung und Forschung (BMBF), Berlin.

Bortz, J. (2005). *Statistik: für Human- und Sozialwissenschaftler* (6. vollst. überarb. u. aktualisierte Aufl.). Heidelberg: Springer.

Bos, W., Bonsen, M., Gröhlich, C., Guill, K., May, P., Rau, A., Stubbe, T.C., Vieluf, U., & Wocken, H. (2007). *KESS 7 – Kompetenzen und Einstellungen von Schülerinnen und Schülern – Jahrgangsstufe 7*. http://www.ifs-dortmund.de/files/KESS-7-Bericht_170309.pdf. Zugegriffen: 14. Mai 2010.

Braun, H., & Wainer, H. (2007). Value-added modeling. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of statistics 26: Psychometrics* (S. 867–892). Boston: Elsevier.

Braun, H., Chudowsky, N., & Koenig, J. (2010). *Getting value out of value-added: Report of a workshop*. Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability; National Research Council.

Briggs, D. C. (2008). *The goals and uses of value-added models*. Paper prepared for a workshop held by the Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation and Educational Accountability sponsored by the National Research Council and the National Academy of Education, Washington, November 13–14, 2008.

Emmrich, R. (2010). *Rückmeldungen VERA 8: Rückmeldeformate und Nutzungsmöglichkeiten Schuljahr 2009/10*. http://www.isq-bb.de/uploads/media/VERA8_2010_Rueckmeldungen_Engl.pdf. Zugegriffen: 14. Mai 2010.

Emmrich, R., Harych, P., Hammer, U., & Hüsemann, D. (2010). *VERA 8: Vergleichsarbeiten in der Jahrgangsstufe 8 im Schuljahr 2008/2009– Länderbericht Brandenburg*. ISQ (Hrsg.). http://www.isq-bb.de/uploads/media/Bericht_Brandenburg_2010_02_15_final.pdf. Zugegriffen: 14. Mai 2010.

Fiege, C. (2007). *Faire Vergleiche in Schulleistungsuntersuchungen und ihre kausaltheoretische Grundlage*. Unveröffentlichte Diplomarbeit, Friedrich-Schiller-Universität Jena.

Fiege, C., Steyer, R., & Nachtigall, C. (2010, Juli). *Which kinds of causal effects are we looking for in educational research? – An application of the theory of causal effects*. Vortrag auf dem Symposium on Causality, Dornburg, Deutschland.

Freie und Hansestadt Hamburg, Behörde für Schule und Berufsbildung, Institut für Bildungsmonitoring. (Hrsg.). (2009). *Bildungsbericht Hamburg 2009*. http://www.bildungsmonitoring.hamburg.de/index.php/file/download/1359. Zugegriffen: 21. Juni 2010.

Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Heidelberg: Springer.

Hartig, J., Klieme, E., & Leutner, D. (Hrsg.). (2008). *Assessment of competencies in educational settings: State of the art and future prospects*. Göttingen: Hogrefe.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlations for planning group-randomized experiments in education. *Educational Evaluation and Policy Analysis, 29,* 60–87.

Helmke, A., & Hosenfeld, I. (2004). Vergleichsarbeiten – Kompetenzmodelle – Standards. In M. Wosnitza, A. Frey, & R. S. Jäger (Hrsg.), *Lernprozesse, Lernumgebungen und Lerndiagnostik. Wissenschaftliche Beiträge zum Lernen im 21. Jahrhundert* (S. 56–75). Landau: Verlag Empirische Pädagogik.

Helmke, A., & Hosenfeld, I. (2005). Standardbasierte Unterrichtsevaluation. In G. Brägger, B. Bucher, & N. Landwehr (Hrsg.), *Schlüsselfragen zur externen Schulevaluation* (S. 127–151). Bern: h.e.p.

Helmke, A., Hosenfeld, I., & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold & C. Griese (Hrsg.), *Schulleitung und Schulentwicklung* (S. 119–144). Hohengehren: Schneider.

Hovestadt, G., & Kessler, N. (2005). 16 Bundesländer – Eine Übersicht zu Bildungsstandards und Evaluationen. In G. Becker, A. Bremerich-Vos, M. Demmer, K. Maag Merki, B. Priebe, K. Schwippert, L. Stäudel, & K. J. Tillmann (Hrsg.), *Standards – Unterrichten zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (Friedrich Jahresheft XXIII 2005, S. 8–10). Seelze: Friedrich.

Isaac, K., & Hosenfeld, I. (2008). Faire Ergebnisrückmeldungen bei Vergleichsarbeiten. In J. Ramseger, & M. Wagener (Hrsg.), *Chancenungleichheit in der Grundschule – Ursachen und Wege aus der Krise* (S. 143–146). Wiesbaden: VS-Verlag für Sozialwissenschaften.

Klieme, E., & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. *Z Pädagogik, 52,* 876–903.

Klieme, E., & Hartig, J. (2008). Kompetenzkonzepte in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. In M. Prenzel, I. Gogolin, & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik* (Sonderheft 8 der Zeitschrift für Erziehungswissenschaft, S. 11–29). Wiesbaden: VS Verlag für Sozialwissenschaften.

KMK (Hrsg.). (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. Bonn: LinkLuchterhand.

Kuper, H., & Schneewind, J. (Hrsg.). (2006). *Rückmeldung und Rezeption von Forschungsergebnissen – Zur Verwendung wissenschaftlichen Wissens im Bildungssystem*. Münster: Waxmann.

Maier, U. (2008). Vergleichsarbeiten im Vergleich – Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessungen in Baden-Württemberg und Thüringen. *Z Erziehungswissensch, 11,* 453–474.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica: RAND Corporation.

Meyer, R. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 16,* 283–301.

Mill, J. S. (1843). Of the four methods of experimental inquiry. *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence, and the methods of scientific investigation* (Bd. 1). London: Longmans, Green, and Co.

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2005). *Zentrale Lernstandserhebungen in Jahrgangsstufe 9 – Schulische Standorttypen und Referenzwerte: Verfahren 2005*. http://www.standardsicherung.schulministerium.nrw.de/lernstand8/upload/download/mat_2005/Standorttypenkonzept_2005.pdf. Zugegriffen: 14. Mai 2010.

Moosbrugger, H., & Kelava, A. (Hrsg.). (2007). *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer.

Müller, A. (2010). *Rückmeldungen nach Vergleichsarbeiten im Kontext des schulischen Qualitätsmanagements. Drei explorative Studien zu Gestaltung und Rezeption im Anschluss an KOALA-S.* Berlin: Mensch und Buch.

Nachtigall, C., & Kröhne, U. (2006). Methodische Anforderungen an schulische Leistungsmessung – Auf dem Weg zu fairen Vergleichen. In H. Kuper, & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen – Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 59–74). Münster: Waxmann.

Nachtigall, C., Kröhne, U., Enders, U., & Steyer, R. (2008). Causal effects and fair comparisons: Considering the influence of context variables on student competencies. In J. Hartig, E. Klieme, & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts: State of the art and future prospects* (S. 315–336). Göttingen: Hogrefe.

Nachtigall, C., Storbeck, I., & Landmann, M. (2009). Belastung oder Chance? Zur Nutzung von Vergleichsarbeiten, Lernstandserhebungen, Kompetenztests, Orientierungsarbeiten und Co. *Schulleitung und Schulentwicklung, 45,* 1–17.

Orth, G. (2002). Vergleichsarbeiten. In H.-G. Rolff, & J. Schmidt (Hrsg.), *Schulaufsicht und Schulleitung in Deutschland*. Neuwied: Luchterhand.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20,* 307–335.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 59–71). Weinheim: Beltz.

Rolff, H.-G. (2002). Rückmeldung und Nutzung der Ergebnisse von großflächigen Leistungsuntersuchungen. Grenzen und Chancen. In R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (S. 75–98). Weinheim: Juventa.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177.

Schrader, F.-W., & Helmke, A. (2008). Determinanten der Schulleistung. In M. Schweer (Hrsg.), *Lehrer-Schüler-Interaktion: Inhaltsfelder, Forschungsperspektiven und methodische Zugänge* (2. Aufl., S. 285–302). Wiesbaden: VS Verlag für Sozialwissenschaften.

Steyer, R., Partchev, I., Kröhne, U., Nagengast, B., & Fiege, C. (in Druck). *Probability and causality*. New York: Springer.

Watermann, R., & Stanat, P. (2004). Schulrückmeldungen in PISA 2000: Sozialnorm- und kriteriumsorientierte Rückmeldeverfahren. *Empirische Pädagogik, 18,* 40–61.

Watermann, R., Stanat, P., Kunter, M., Klieme, E., & Baumert, J. (2003). Schulrückmeldungen im Rahmen von Schulleistungsuntersuchungen: Das Disseminationskonzept von PISA-2000. *Z Pädagogik, 49,* 92–111.

Wegscheider, K. (2004). Methodische Anforderungen an Einrichtungsvergleiche (‚Profiling') im Gesundheitswesen. *Z Ärztliche Fortbildung Qualität Gesundheitswesen, 98,* 647–654.

Weinert, F. E. (Hrsg.). (2002). *Leistungsmessungen an Schulen*. Weinheim: Beltz.