Dirk Richter/Katrin Böhme/Michael Becker/Hans Anand Pant/Petra Stanat

Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten

Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern

Zusammenfassung: Die Vergleichsarbeiten (VERA) sind seit mehreren Jahren ein wichtiges Instrument der Kompetenzdiagnostik, das auf den Bildungsstandards der Kultusministerkonferenz basiert. Sie dienen in erster Linie der Unterrichts- und Schulentwicklung, werden teilweise aber auch zur flächendeckenden Information der Schulaufsicht über den Leistungsstand von Einzelschulen genutzt. Der vorliegende Beitrag untersucht, inwieweit diese Funktionen von Lehrkräften wahrgenommen werden und in welcher Beziehung sie zum Unterricht der Lehrkräfte und den Kompetenzen der Schülerinnen und Schüler stehen. Die Studie basiert auf Daten des IQB-Ländervergleichs 2011 in der Primarstufe, in dem Kompetenzen in den Fächern Deutsch und Mathematik erhoben wurden. Die Analysen zeigen, dass Lehrkräfte, die VERA als Mittel der Unterrichtsentwicklung begreifen, ihren Unterricht verstärkt auf die Entwicklung von Kompetenzen ausrichten und eine stärkere Differenzierung im Unterricht vornehmen. Weiterhin erreichen Schülerinnen und Schüler von Lehrkräften mit diesen Überzeugungen bessere Ergebnisse im Lesen und in Mathematik, auch nach Berücksichtigung individueller und klassenbezogener Hintergrundmerkmale.

Schlagworte: Funktionen von Schulleistungstests, Grundschule, Ländervergleich, Lesekompetenz, mathematische Kompetenzen

1. Einleitung

Nach dem erwartungswidrig schlechten Abschneiden Deutschlands in den Schulleistungsstudien TIMSS (Baumert et al., 1997) und PISA (Baumert et al., 2001) setzte im deutschen Bildungswesen ein umfassender Reformprozess ein, der darauf abzielte, das input-orientierte Steuerungsmodell durch Elemente einer output-orientierten Steuerung zu ergänzen (z. B. Altrichter & Maag-Merki, 2010). Eine Grundlage dieser Optimierungsbemühungen der Kultusministerkonferenz (KMK) bildet die Einführung länderübergreifend verbindlicher Bildungsstandards, die beschreiben, welche Kompetenzen Schülerinnen und Schüler bis zum Ende eines Bildungsabschnitts im Regelfall erworben haben sollen (KMK, 2005). Eine Möglichkeit zur output-orientierten Steuerung besteht darin, Informationen über schulische Leistungserträge zu sammeln und Akteuren des Schulsystems hierüber Rückmeldung zu geben. Zu diesem Zweck hat die KMK eine Gesamtstrategie zum Bildungsmonitoring verabschiedet (KMK, 2006). Ein Bestandteil dieser Strategie ist die regelmäßige und flächendeckende Durchführung von Untersuchungen des Leistungsstandes von Schülerinnen und Schülern eines Jahrgangs

in allen Schulen und Klassen. Diese Erhebungen, die eng auf die fachlichen Zielvorgaben der Bildungsstandards bezogen sind, werden in der Regel als „Vergleichsarbeiten“[1] (VERA), in einzelnen Ländern aber auch als „Lernstandserhebungen“ oder „Kompetenztests“ bezeichnet.

Die primäre Funktion von VERA besteht darin, Prozesse der Schul- und Unterrichtsentwicklung durch Feedback über den Leistungsstand von Schulklassen zu unterstützen (KMK, 2010, 2012). Ein solcher Entwicklungsprozess setzt voraus, dass Lehrkräfte die Ergebnisse des Leistungstests zunächst analysieren, anschließend angemessene Maßnahmen zur Weiterentwicklung des eigenen Unterrichts ableiten, diese umsetzen und laufend evaluieren (vgl. Helmke, 2004). Eine weitere Funktion von VERA umfasst die Unterstützung der Schulaufsicht und/oder der Schulinspektion, die in einigen Ländern Einsicht in die VERA-Ergebnisse auf Schul- und Klassenebene erhält (KMK, 2012). Somit nimmt VERA – je nach Landesregelung – eine Doppelfunktion ein: Die Tests dienen einerseits der Schul- und Unterrichtsentwicklung, vermittelt über das Feedback von Schülerleistungen auf der Ebene der Klasse oder Schule, und andererseits der Rechenschaftslegung in Bezug auf Schülerleistungen auf der Ebene einzelner Klassen, Schulen, Schularten, Regionen oder ein Bundesland.

Welche dieser beiden Funktionen Lehrkräfte VERA zuschreiben, ist bislang in der empirischen Forschung kaum betrachtet worden. Es gibt jedoch Hinweise darauf, dass die von Lehrkräften wahrgenommenen Funktionen von VERA nicht immer deckungsgleich mit den intendierten Funktionen dieses Instruments sind (Kühle & Peek, 2007; Kuper & Hartung, 2007). Da die Mehrzahl der empirischen Arbeiten zu diesem Thema aus den Anfangsjahren der Vergleichsarbeiten stammt, ist unklar, ob Lehrkräfte gut zehn Jahre nach der Einführung von VERA nach wie vor ähnliche Überzeugungen aufweisen. Auch lässt sich aus den bisher vorliegenden Arbeiten nicht ableiten, ob die Überzeugungen der Lehrkräfte relevant für den Umgang mit Leistungsrückmeldungen und die Unterrichtsgestaltung sind.

Ziel der vorliegenden Arbeit ist es zunächst, Überzeugungen von Lehrkräften zur Entwicklungs- und Kontrollfunktion von VERA zu beschreiben. Anschließend wird empirisch überprüft, in welcher Beziehung diese Überzeugungen sowohl zu Aspekten der Unterrichtsgestaltung als auch zu den von Schülerinnen und Schülern erreichten Kompetenzständen stehen. Zur Einbettung dieser Analysen werden zuerst die historische Entwicklung sowie die intendierten Funktionen der Vergleichsarbeiten beschrieben. Anschließend wird dargestellt, welche Folgen aus der Einführung von Schulleistungsstudien für das professionelle Handeln von Lehrkräften und die Kompetenzen von Schülerinnen und Schülern resultieren können. Der Fokus liegt dabei auf den Überzeugungen bezüglich der Vergleichsarbeiten in der Primarstufe, VERA-3.

1 In diesem Artikel werden nachfolgend ausschließlich die Begriffe Vergleichsarbeiten bzw. VERA verwendet, auch wenn in einzelnen Ländern die Begriffe für die jahrgangsbezogenen Tests in Jahrgangsstufe 3 und 8 abweichen.

1.1 Vergleichsarbeiten in Deutschland: Historische Entwicklung und Ziele der VERA-Tests

Die Vergleichsarbeiten im Primarbereich wurden erstmals 2003 in Rheinland-Pfalz im Fach Mathematik durch die Universität Koblenz-Landau durchgeführt (Helmke & Hosenfeld, 2003). Im Jahr 2004 schlossen sich die Länder Berlin, Brandenburg, Bremen, Mecklenburg-Vorpommern, Nordrhein-Westfalen und Schleswig-Holstein an, und VERA wurde um das Fach Deutsch ergänzt. Die Testdurchführung erfolgte jährlich kurz nach Beginn des Schuljahres in allen Grundschulklassen der vierten Jahrgangsstufe. Seit dem Schuljahr 2007/2008 werden die Vergleichsarbeiten in der dritten Jahrgangsstufe geschrieben (VERA-3). Dadurch erhalten Lehrkräfte bereits ein Jahr vor dem Übergang in die weiterführenden Schulen eine externe Rückmeldung zum Leistungsstand ihrer Schülerinnen und Schüler und können ggf. zusätzliche Förderangebote bereitstellen. Seit dem Schuljahr 2008/2009 beteiligen sich alle 16 Länder in der Bundesrepublik Deutschland an VERA-3, und auch Südtirol sowie die deutschsprachige Gemeinschaft Belgiens haben sich angeschlossen.

Mit der Durchführung von VERA-3 im Jahr 2010 ist die Entwicklung der Testaufgaben sowie der begleitenden didaktischen Materialien an das Institut zur Qualitätsentwicklung im Bildungswesen (IQB) übergegangen, um eine engere Anbindung an die länderübergreifenden Bildungsstandards zu etablieren. In jährlichen, groß angelegten Pilotierungsstudien werden die Aufgaben des kommenden VERA-Durchgangs einer repräsentativen Stichprobe von Schülerinnen und Schülern gemeinsam mit bereits normierten Aufgaben zu den Bildungsstandards vorgelegt. Auf der Basis dieser gemeinsamen Datenerhebung wird eine psychometrische Anbindung von VERA an die Metrik der Bildungsstandards ermöglicht. Seit dem Schuljahr 2007/2008 werden in den Fächern Deutsch, Mathematik sowie in der ersten Fremdsprache (Englisch und Französisch) auch Vergleichsarbeiten in der achten Jahrgangsstufe durchgeführt (VERA-8). Einige Länder lassen zudem Vergleichsarbeiten in der sechsten Jahrgangsstufe schreiben.

VERA ist derzeit als Vollerhebung in allen Klassen der jeweiligen Jahrgangsstufe (drei bzw. acht) angelegt, d. h. die Untersuchung des Kompetenzstandes erfolgt in allen Schulen und Klassen innerhalb eines Bundeslandes (KMK, 2012). In der Regel obliegt die Testdurchführung ebenso wie die Auswertung der Schülerantworten der Lehrkraft der getesteten Klasse. Aus diesem Grund ist der Grad der Standardisierung in VERA wesentlich geringer als in internationalen Schulleistungsstudien oder den Ländervergleichen des IQB. Die Ergebnisse der VERA-Tests werden den Lehrkräften sowie ihren Schülerinnen und Schülern bzw. deren Eltern zurückgemeldet und – je nach Bundesland – auch der Schulleitung, der Schulaufsicht und der Schulinspektion zur Verfügung gestellt. Darüber hinausgehende Veröffentlichungen der Ergebnisse von Einzelschulen, etwa in Form von Rankings, finden nicht statt (KMK, 2012).

Den Vergleichsarbeiten werden zwei wesentliche Funktionen zugeschrieben: zum einen die der Unterstützung von *Schul- und Unterrichtsentwicklung,* zum anderen die der *Rechenschaftslegung* (vgl. Maier, Metz, Bohl, Kleinknecht & Schymala, 2012).

Welche dieser Funktionen VERA aus Sicht von Lehrkräften erfüllt, wurde bislang in nur wenigen Arbeiten empirisch untersucht. In einer qualitativen Studie konnten Kuper und Hartung (2007) zeigen, dass sich Lehrkräfte darin unterscheiden, ob sie die Ergebnisse zur Reflexion der eigenen Unterrichtspraxis nutzen oder sie als eine fremdbestimmte Kontrolle der eigenen Arbeit erleben. In einer quantitativen Erhebung differenzierten Kühle und Peek (2007) drei Funktionen von Vergleichsarbeiten: Individualdiagnose, Unterrichtsentwicklung und Systemmonitoring. Sie konnten zeigen, dass die erfragten Funktionen ähnlich bewertet werden und im Durchschnitt keine stark befürwortet oder abgelehnt wird. Da die Daten, die diesen Studien zugrunde lagen, in den ersten Jahren von VERA erhoben wurden, lässt sich vermuten, dass die mehrjährige Erfahrung mit dem Instrument die Wahrnehmung bei Lehrkräften verändert hat. Die Betonung der Unterrichts- und Schulentwicklungsfunktion von VERA durch politische Entscheidungsträger (vgl. KMK, 2006, 2012, 2013) könnte dazu beigetragen haben, dass auch Lehrkräfte den Test heute anders wahrnehmen und möglicherweise verstärkt im Sinne der Unterrichts- und Schulentwicklungsfunktion interpretieren. Empirische Erkenntnisse liegen hierzu jedoch nicht vor.

1.2 Konsequenzen von Schulleistungstests für den Unterricht und die Kompetenzen von Schülerinnen und Schülern

Die Einführung von flächendeckenden Vergleichsarbeiten wurde vonseiten der Bildungspolitik mit der Erwartung verbunden, dass die Auseinandersetzung mit den zurückgemeldeten Testergebnissen zu einer Verbesserung der Unterrichtsqualität und der Lernergebnisse beiträgt (KMK, 2010, 2012). Den Prozess der Auseinandersetzung beschrieb Helmke (2004) durch ein theoretisches Rahmenmodell, welches vier aufeinanderfolgende Schritte vereint. Der Verarbeitungsprozess beginnt mit der *Rezeption* der Ergebnisse durch die Lehrkraft und setzt sich fort in der Analyse möglicher Ursachen *(Reflexion)*. Anschließend werden auf dieser Grundlage Maßnahmen zur Optimierung des Unterrichts ergriffen *(Aktion),* die abschließend auf ihre Wirksamkeit hin geprüft werden *(Evaluation)*. Dieser idealtypische Verlauf beschreibt ein Verfahren zur qualitativen Weiterentwicklung des Unterrichts, die sich schließlich auch in den Ergebnissen von Schulleistungstests niederschlagen sollte.

Empirische Studien haben gezeigt, dass die überwiegende Mehrheit der Lehrkräfte, deren Klassen in VERA getestet wurden, die schriftlichen Ergebnisrückmeldungen zur Kenntnis nimmt (Dedering, 2011). Die Voraussetzung für die weiteren Schritte innerhalb der von Helmke (2004) beschriebenen Wirkungskette *(Rezeption)* ist somit gegeben. Jedoch zieht nach bisherigen empirischen Erkenntnissen nur ein geringer Teil von Lehrkräften die Rückmeldungen zu den Vergleichsarbeiten zur Ableitung von Maßnahmen für den eigenen Unterricht heran *(Reflexion und Aktion)* (Bonsen, Büchter & Peek, 2006; Dedering, 2011; Groß Ophoff, Koch, Hosenfeld & Helmke, 2006; Kühle & Peek, 2007; Maier, 2007). Mögliche Unterrichtsmodifikationen *(Aktion)* beziehen sich vor allem auf die Verwendung von Aufgaben aus dem Test, eine veränderte Unterrichts-

gestaltung (z. B. bei der Leistungsdifferenzierung) sowie eine Vertiefung und Wiederholung bestimmter Inhaltsbereiche (Groß Ophoff et al., 2006; Koch, Groß Ophoff, Hosenfeld & Helmke, 2006; Kühle & Peek, 2007; Schneewind & Kuper, 2009).

Aus Schulleistungstests, die primär der Rechenschaftslegung gegenüber der Bildungsverwaltung dienen, werden mitunter auch andere Konsequenzen für den Unterricht abgeleitet *(Aktion)*. Entsprechend eines von Donald T. Campbell (1979) formulierten Gesetzes führt ein Testsystem, in dem Schülerleistungen die Grundlage von Entscheidungsprozessen sind, dazu, dass diese möglicherweise Verfälschungstendenzen unterliegen. Wird also ein Leistungstest zur Kontrolle schulischer Ergebnisse verwendet und werden darüber hinaus mit den Testergebnissen Konsequenzen für die schulischen Akteure verknüpft, so ist zu erwarten, dass sich das Handeln der Lehrkräfte auf die Optimierung der Testergebnisse ausrichtet. Dies umfasst einerseits intendierte Folgen wie die Ausweitung von Lernzeit und die Verbesserung von Unterrichtsqualität, andererseits aber auch nicht-intendierte Folgen wie die Einschränkung des Curriculums und die Reduzierung der Unterrichtszeit nicht getesteter Fächer (*Aktion* im Sinne von Helmke, 2004).

In den Vereinigten Staaten gibt es umfangreiche Forschung zu den Effekten von Schulleistungstests, die primär der Kontrolle schulischer Leistungen dienen (Bellmann & Weiß, 2009; Hamilton et al., 2007; Maag-Merki, 2010; Maier & Kuper, 2012). Übereinstimmend zeigen mehrere Studien, dass sich die Unterrichtszeit zugunsten der geprüften Fächer (Englisch und Mathematik) deutlich erhöhte (Griffith & Scharmann, 2008; Hamilton et al., 2007; McMurrer, 2007; Rentner et al., 2006; Smith & Kovacs, 2011). Darüber hinaus liegen Belege für eine stärkere Verengung des Curriculums auf die im Test bzw. in den Standards enthaltenen Themen vor (Hamilton et al., 2007; McMurrer, 2007; Smith & Kovacs, 2011). Weiterhin kam es in den USA auch zu Angleichungsprozessen bei den im Unterricht und im Test verwendeten Prüfungsformaten (Hamilton et al., 2007). Auch finden sich Hinweise, dass die Einführung von Leistungstests zur Rechenschaftslegung in den USA im Rahmen von *No Child Left Behind* zu einer Erhöhung des Anteils der als *proficient* klassifizierten Schüler beigetragen hat (Kober, Chudowsky & Chudowsky, 2008). Da sich dieser Leistungszuwachs jedoch nicht in gleichem Maße im stichprobenbasierten *National Assessment of Educational Progress* nachweisen ließ (Kober et al., 2008; Lee, 2007), muss diese vermeintlich positive Leistungsentwicklung *sensu* Campbell eher als Artefakt und Konsequenz eines auf Rechenschaftslegung ausgerichteten Testsystems interpretiert werden.

Die Untersuchung möglicher Konsequenzen von Leistungstests für den Unterricht und die Kompetenzentwicklung von Schülerinnen und Schülern setzt somit voraus, auch die Funktionen, die ein Test in einem Bildungssystem erfüllt, zu berücksichtigen. Dies ist erforderlich, da Tests mit verschiedenen Funktionen bei Lehrkräften zu unterschiedlichen Konsequenzen in der Reflexion und in der Unterrichtsgestaltung, also im Schritt der *Aktion* führen können. Für die deutschen Vergleichsarbeiten bedeutet dies, dass die Doppelfunktion von Schul- und Unterrichtsentwicklung einerseits sowie Rechenschaftslegung andererseits neben positiven Effekten auf die Unterrichtskultur möglicherweise auch negative Folgen mit sich bringen kann. Bislang gibt es nur wenige

Studien, die der Vermutung bezüglich potenziell negativer Folgen von Vergleichsarbeiten empirisch nachgegangen sind. Erste Ergebnisse einer baden-württembergischen Lehrkräftebefragung weisen nicht darauf hin, dass der Test zu einer starken Verengung des Curriculums beiträgt (Wacker & Kramer, 2012). Bei dieser Erhebung ist jedoch unberücksichtigt geblieben, dass Lehrkräfte unterschiedliche Funktionen mit dem Test assoziieren. Es bleibt somit die Frage offen, inwiefern die subjektiv wahrgenommene Funktion der Tests mit möglichen positiven und negativen Folgen für den Unterricht verbunden ist.

1.3 Fragestellung der Studie

Die vorliegende Arbeit untersucht die Wahrnehmung der den Vergleichsarbeiten zugeschriebenen Funktionen *Unterrichtsentwicklung* sowie *Rechenschaftslegung* bzw. *Kontrolle von Schulen.* Der erste Teil der Studie betrachtet Zusammenhänge zwischen diesen Überzeugungen und Merkmalen des Unterrichts. Dabei werden zum einen Merkmale untersucht, die als positive Folgen von VERA gelten *(stärkere Kompetenzorientierung* und *Differenzierung),* zum anderen werden auch ambivalente oder negative Folgen *(Verengung des Lehrplans* und *keine Veränderung)* in den Blick genommen. Es lässt sich vermuten, dass Lehrkräfte, die Vergleichsarbeiten als Instrument der Unterrichtsentwicklung begreifen, eine stärkere Kompetenzorientierung und Differenzierung in ihrem Unterricht aufweisen. Bei Lehrkräften, die VERA als Kontrollinstrument bewerten, ist zu erwarten, dass sie eher eine Verengung des Lehrplans vorgenommen haben.

Im zweiten Teil der Studie werden Zusammenhänge zwischen den Überzeugungen der Lehrkräfte und den erreichten Schülerkompetenzen im Lesen und in Mathematik untersucht. Anknüpfend an die Hypothese zur Bedeutung der Überzeugungen für den Unterricht lässt sich vermuten, dass Schülerinnen und Schüler von Lehrkräften, die VERA als Instrument zur Unterrichtsentwicklung begreifen, bessere Leistungen in den betrachteten Kompetenzbereichen erzielen. Für Lehrkräfte, die VERA als Kontrollinstrument auffassen, sollte sich ebenfalls ein positiver Zusammenhang zu den schülerseitigen Kompetenzständen ergeben, da Lehrkräfte mit entsprechender Überzeugung darum bemüht sein sollten, dass ihre Schülerinnen und Schüler – auch in schulinternen Vergleichen – möglichst gut abschneiden.

2. Methode

2.1 Studie und Stichprobe

Die vorliegende Studie basiert auf den Daten des IQB-Ländervergleichs 2011 im Primarbereich, der anhand einer bundesweit repräsentativen Stichprobe Kompetenzen von Schülerinnen und Schülern der vierten Jahrgangsstufe überprüfte (Stanat, Pant, Böhme & Richter, 2012). Eingesetzt wurden Leistungstests im Fach Deutsch zu den Bereichen

Lesen und Zuhören sowie im Fach Mathematik. Weiterhin beinhaltete die Studie eine Befragung der Deutsch- und Mathematiklehrkräfte der beteiligten Klassen zu ihrer Aus- und Fortbildung, zur Unterrichtsgestaltung sowie zur wahrgenommenen Funktion der Vergleichsarbeiten. Insgesamt nahmen an der Untersuchung Schülerinnen und Schüler einer zufällig ausgewählten Klasse von 1295 Grundschulen, 3 Walddorfschulen und 51 Förderschulen teil (Richter et al., 2012). Die Analysen für die vorliegende Studie beschränken sich auf Grund- und Waldorfschulen; somit reduziert sich die Stichprobe in dieser Arbeit auf 1298 Schulen und die darin getesteten Schülerinnen und Schüler sowie ihre Lehrkräfte.

Die Lehrerstichprobe umfasst insgesamt 1757 Grundschullehrkräfte, von denen 567 in der getesteten Klasse nur das Fach Deutsch, 545 nur das Fach Mathematik und 611 beide Fächer unterrichteten. Von 34 Lehrkräften liegen keine Angaben zum unterrichteten Fach in der Klasse vor. Die befragten Lehrkräfte sind im Durchschnitt 47.7 Jahre alt (SD = 10.3) und überwiegend weiblich (88.3 %).

Zur Vorhersage der Testleistungen werden die Schülerinnen und Schüler jeweils ihrer Deutsch- bzw. Mathematiklehrkraft zugeordnet. Im Fach Deutsch liegen für insgesamt 22 389 von 26 029 Schülerinnen und Schülern Fragebögen von Lehrkräften vor. Dies entspricht einem Anteil von 86.0 %. Die Kinder waren zum Zeitpunkt der Erhebung im Durchschnitt 10.5 Jahre alt (SD = 0.5 Jahre) und 49.6 % von ihnen waren weiblich. Im Fach Mathematik konnten den teilnehmenden Lehrkräften insgesamt 22 002 von 26 016 Schülerinnen und Schülern zugeordnet werden, was einer Quote von 84.6 % entspricht. Die Kinder waren ebenfalls im Durchschnitt 10.5 Jahre alt (SD = 0.5 Jahre) und 49.4 % von ihnen waren weiblich. Alle Klassen, für die keine Angabe einer Lehrkraft vorlag, wurden von den Analysen ausgeschlossen. Im Fach Deutsch betrifft dies 178 Klassen und im Fach Mathematik 195 Klassen.

2.2 Instrumente

Überzeugungen zu VERA (Lehrerebene). Die Überzeugungen der Lehrkräfte zu den Funktionen von VERA wurden mit zwei Skalen erhoben. Die erste Skala erfasst, inwiefern der Test aus Sicht der Lehrkräfte ein diagnostisches Instrument darstellt, dessen Ergebnisse zur Unterrichtsentwicklung genutzt werden können. Die zweite Skala beschreibt die Überzeugung, dass VERA ein Instrument zur Kontrolle von Lehrkräften und Schulen ist. Die Items wurden am IQB entwickelt und erstmals im Ländervergleich 2011 eingesetzt. Die Skalen repräsentieren zwei Dimensionen, die schwach positiv miteinander korreliert sind (r = .13, p < .05). Die Reliabilitäten der Skalen, die Anzahl der zugrunde liegenden Items und Beispielitems sind in Tabelle 1 aufgeführt. Alle Items wurden von der Lehrkraft auf einer vierstufigen Likert-Skala von (1) *stimme nicht zu* bis (4) *stimme völlig zu* eingeschätzt.

Veränderung des Unterrichts (Lehrerebene). Weiterhin wurden die Lehrkräfte zu Veränderungen befragt, die sie in ihrem eigenen Unterricht infolge der Einführung von Leistungsvergleichen wahrgenommen haben. Die Liste der abgefragten Veränderungen

Skala	Itemzahl	Reliabilität Cronbachs Alpha	Beispielitem
Überzeugungen zu den Funktionen von VERA			
Unterrichtsentwicklung	7	.88	Die Ergebnisse der landesweiten Lernstandserhebungen/Vergleichsarbeiten (VERA-3) geben wichtige Anhaltspunkte darauf, welche Kompetenzen noch stärker gefördert werden müssen.
Kontrolle	3	.78	Die Ergebnisse der landesweiten Lernstandserhebungen/Vergleichsarbeiten (VERA-3) dienen dazu, die Schulaufsichtsbehörden über die Leistungen von Schulen zu informieren.
Veränderungen im Unterricht			
Kompetenzorientierung	4	.80	Ich konzentriere mich stärker auf die Bildungsstandards der Kultusministerkonferenz.
Differenzierung	2	.87	Ich konzentriere mich stärker auf Schülerinnen und Schüler am unteren Ende des Leistungsspektrums.
Verengung des Lehrplans	4	.77	Ich nehme mir weniger Freiheiten in der inhaltlichen Gestaltung meines Unterrichts.
Keine Veränderung	2	.62	Ich halte es für falsch, wegen Leistungsvergleichen Veränderungen in meinem Unterricht vorzunehmen.

Tab. 1: Übersicht der eingesetzten Skalen mit Angaben zu Reliabilität und Beispielitems

wurde auf Grundlage einer Studie von Hamilton et al. (2007) entwickelt und von uns für die Verwendung im deutschen Kontext angepasst. Abgefragt wurden Verhaltensweisen, die sich zu vier Skalen zusammenfassen lassen, die ebenfalls in Tabelle 1 aufgeführt sind. Die Items dieser Skalen wurden auf einer Likert-Skala von (1) *trifft nicht zu* bis (4) *trifft zu* eingeschätzt. Hierbei repräsentieren die Konstrukte *Kompetenzorientierung* und *Differenzierung* positiv zu bewertende Folgen von VERA, die beiden Skalen, *Verengung des Lehrplans* und *keine Veränderung,* hingegen ambivalente bzw. negative Folgen.

Leistungstests (Schülerebene). Weiterhin wurden Ergebnisse der Kompetenztests des Ländervergleichs 2011 im Lesen und in Mathematik in die Analysen einbezogen. Zur Erfassung der Lesekompetenz beantworteten Schülerinnen und Schüler Fragen, die sich auf Sachtexte und literarische Texte bezogen (Böhme & Bremerich-Vos, 2012). In Abhängigkeit vom jeweiligen Testheft betrug die Testzeit für die Erfassung der Lesekompetenz 20 oder 40 Minuten. Über alle Testheftversionen hinweg kamen 11 Aufgaben (Stimulustexte) mit insgesamt 80 Items zum Einsatz (Weirich, Haag & Roppelt, 2012).

Der Test zur Erfassung mathematischer Kompetenzen umfasste Aufgaben zu allen fünf inhaltlichen Kompetenzbereichen (Leitideen), die in den Bildungsstandards für die

vierte Jahrgangsstufe beschrieben werden: (1) Zahlen und Operationen, (2) Raum und Form, (3) Muster und Strukturen, (4) Größen und Messen sowie (5) Daten, Häufigkeit und Wahrscheinlichkeit (Roppelt & Reiss, 2012). Für die Analysen der vorliegenden Arbeit wurde die im IQB-Ländervergleich 2011 berichtete Globalskala mathematischer Kompetenz genutzt, welche alle getesteten Leitideen integriert. Wie im Fach Deutsch standen den Schülerinnen und Schülern insgesamt 80 Minuten zur Bearbeitung des Tests zur Verfügung. Berücksichtigt man die Aufgaben aller eingesetzten Testheftversionen, kamen insgesamt 201 Aufgaben mit 330 Items zum Einsatz (Weirich et al., 2012). Zur Schätzung der Leistungswerte wurde, entsprechend dem üblichen Vorgehen in Large-Scale-Assessments, die *Plausible-Value*-Methode verwendet (von Davier, Gonzalez & Mislevy, 2009). Die Skalen beider Kompetenzbereiche wurden jeweils auf einen Mittelwert von 500 Punkten und eine Standardabweichung von 100 Punkten normiert.

Als Kontrollvariablen wurden auf Lehrerebene das Geschlecht der Lehrkraft sowie die Berufserfahrung berücksichtigt. Auf Schülerebene wurden darüber hinaus die zuhause gesprochene Sprache und der sozio-ökonomische Status der Eltern, gemessen durch den HISEI (Highest International Socio-Economic Index; Ganzeboom, de Graaf, Treiman & de Leeuw, 1992), als Kontrollvariablen verwendet.

2.3 Analysen

Der Zusammenhang zwischen den Überzeugungen der Lehrkräfte und den berichteten Veränderungen im eigenen Unterricht wurde mit einer Regressionsanalyse untersucht, in die Hintergrundmerkmale der Lehrkräfte und der unterrichteten Klassen als Kontrollvariablen eingingen. Die Vorhersage der Kompetenzen von Schülerinnen und Schülern im Lesen und in Mathematik erfolgte mit Mehrebenenmodellen (Raudenbush & Bryk, 2002). In diesen Modellen wird berücksichtigt, dass Schülerinnen und Schüler in Klassen gruppiert sind und somit keine vollständige Unabhängigkeit der Beobachtungen gegeben ist. Bei den Schätzungen der Modellparameter und ihrer Standardfehler wird diese hierarchische Struktur einbezogen. Sowohl in die Regressions- als auch in die Mehrebenenmodelle gehen die Überzeugungen der Lehrkräfte als messfehlerbereinigte latente Variablen ein, alle anderen Prädiktoren sowie die abhängigen Variablen werden als manifeste Variablen modelliert. Sämtliche Analysen wurden in dem Programm M*plus* 7 mit dem *Full-Information-Maximum-Likelihood*-Schätzer durchgeführt (Muthén & Muthén, 1998–2012). Die Ergebnisse der Regressions- und Mehrebenenanalysen wurden als unstandardisierte Koeffizienten berichtet und interpretiert, wobei die Überzeugungen der Lehrkräfte als standardisierte Variablen eingehen ($M = 0$, $SD = 1$).

3. Ergebnisse

Zunächst werden die deskriptiven Ergebnisse für die Fragebogenskalen und Kompetenztests dargestellt, um die Verteilungen dieser Variablen zu beschreiben (vgl. Tabelle 2). Neben den statistischen Kennwerten der Verteilungen werden die Ergebnisse von *t*-Tests berichtet, mit denen geprüft wird, ob die Mittelwerte der Fragebogenskalen vom theoretischen Mittel abweichen (H_0: $\mu = 2.5$). Für die Skalen *Unterrichtsentwicklung* und *Kontrolle* liegt der Mittelwert statistisch signifikant unter dem theoretischen Mittelwert, jedoch befinden sich beide Skalenwerte noch in der Mitte des Wertebereichs. Somit lässt sich im Durchschnitt keine dominante Wahrnehmung einer der beiden Funktionen erkennen. Die Mittelwerte der Skalen, die die Veränderungen im Unterricht abbilden, unterscheiden sich ebenfalls signifikant vom theoretischen Mittel der Skala. Für die *Kompetenzorientierung,* die *Differenzierung* und die *Verengung des Lehrplans* zeigen sich insgesamt niedrigere Mittelwerte, wobei lediglich die beiden letztgenannten Konstrukte praktisch bedeutsame Abweichungen vom Mittelwert der Skala aufweisen. Für die Skala *keine Veränderung* zeigt sich ein vergleichsweise hoher Mittelwert, der statistisch signifikant vom theoretischen Mittelwert der Skala abweicht. Im Durchschnitt berichten Lehrkräfte also eher, keine Veränderungen infolge von Leistungstests vorgenommen zu haben. Für jede Skala zeigt sich darüber hinaus eine deutliche Streuung in den Ausprägungen, was darauf schließen lässt, dass Lehrkräfte die mit VERA verbundenen Funktionen unterschiedlich wahrnehmen und in unterschiedlichem Maße Veränderungen im eigenen Unterricht berichten.

Die Ergebnisse der Kompetenztests im Lesen und in Mathematik unterscheiden sich von den Kennwerten der Gesamtstichprobe des Ländervergleichs 2011 ($M = 500$, $SD = 100$), da in die hier berichtete Analyse nur Schülerinnen und Schüler von Regelschulen eingehen, denen eine Lehrkraft zugeordnet werden konnte. In dieser Teilstichprobe fallen deshalb die Mittelwerte etwas höher und die Standardabweichungen etwas geringer aus als in der Gesamtgruppe.

Die Zusammenhänge zwischen den Überzeugungen der Lehrkräfte zu den Funktionen von VERA und den berichteten Veränderungen im Unterricht wurden mit Regressionsanalysen auf Klassenebene untersucht (vgl. Tabelle 3). Die Modelle 1 und 2 beziehen sich auf positive, wünschenswerte unterrichtsbezogene Veränderungen *(Kompetenzorientierung* und *Differenzierung),* während die Modelle 3 und 4 ambivalente/negative Folgen *(Verengung des Lehrplans* und *keine Veränderung)* in den Blick nehmen. Als Kontrollvariablen dienen das Geschlecht und die Berufserfahrung der Lehrkräfte sowie auf Klassenebene aggregierte Angaben zur Familiensprache und zum sozio-ökonomischen Status der Schülerinnen und Schüler.

In *Modell 1* zeigt sich unter Kontrolle aller aufgeführten Kovariaten ein signifikant positiver Zusammenhang zwischen der wahrgenommenen Unterrichtsentwicklungsfunktion von VERA und der Kompetenzorientierung im eigenen Unterricht ($b = 0.52$, $p < .05$). In *Modell 2* besteht ebenfalls ein signifikant positiver Zusammenhang zwischen der wahrgenommenen Unterrichtsentwicklungsfunktion und der berichteten Differenzierung ($b = 0.38$, $p < .05$), also einer intensiveren und passgenauen Förderung

Variablen	*N*	*M*	*SD*	Schiefe	*t*-Test: μ = 2.5 *p*-Wert
Überzeugungen zu den Funktionen von VERA					
Unterrichtsentwicklung	1 691	2.42	0.64	−0.30	< .01
Kontrolle	1 659	2.44	0.73	−0.19	< .01
Veränderungen im Unterricht					
Kompetenzorientierung	1 657	2.42	0.68	−0.53	< .01
Differenzierung	1 649	2.20	0.76	0.02	< .01
Verengung des Lehrplans	1 662	1.70	0.57	0.52	< .01
Keine Veränderung	1 574	2.71	0.76	−0.25	< .01
Kompetenztests					
Lesekompetenz	22 389	504.87	95.89	−0.17	–
Mathematische Kompetenzen	22 002	505.08	96.10	−0.11	–

Anmerkungen. Die Kennwerte für die Leistungstests basieren auf gewichteten Angaben und beziehen sich auf Schülerinnen und Schüler, denen Lehrkräfte zugeordnet werden konnten.

Tab. 2: Mittelwerte, Standardabweichungen und Schiefe für die verwendeten Fragebogenskalen und Kompetenzmaße

von leistungsschwachen und leistungsstarken Schülergruppen. Im Gegensatz zur wahrgenommenen Unterrichtsentwicklungsfunktion zeigen die Modelle 1 und 2 aber keinen Zusammenhang zwischen der Wahrnehmung von VERA als Kontrollinstrument und den hier betrachteten positiven Outcomes, also Kompetenzorientierung und Differenzierung.

Die in *Modell 3* untersuchte Verengung des Lehrplans lässt sich sowohl durch die Wahrnehmung von VERA als Instrument zur Unterrichtsentwicklung ($b = 0.13$, $p < .05$) als auch durch die Wahrnehmung als Kontrollinstrument ($b = 0.12$, $p < .05$) vorhersagen. Eine Verengung des Lehrplans tritt also häufiger auf, wenn VERA stärker als Kontrollinstrument wahrgenommen wird, allerdings findet sich ebenso ein positiver Zusammenhang für die Interpretation von VERA als Unterrichtsentwicklungsinstrument. In *Modell 4* ist ein negativer Zusammenhang zwischen der wahrgenommenen Unterrichtsentwicklungsfunktion und der berichteten Konstanz in der Unterrichtsgestaltung zu verzeichnen ($b = -0.33$, $p < .05$). Demnach nehmen Lehrkräfte nach eigenen Angaben häufiger Veränderungen im Unterricht vor, wenn VERA aus ihrer Sicht zur Unterrichtsentwicklung dient und diagnostische Informationen bereitstellt.

In allen vier Modellen ist der Einfluss der zusätzlich berücksichtigten Kontrollvariablen gering. Lediglich in Modell 4, das das Ausbleiben von Veränderungen in der Unterrichtsgestaltung untersucht, zeigt auch das Geschlecht der Lehrkraft einen Effekt, und zwar in dem Sinne, dass vor allem Frauen dazu neigen, in ihrem Unterricht keine Veränderungen vorzunehmen ($b = 0.20$, $p < .05$). Auch ein höherer sozio-ökonomischer

Prädiktoren	Modell 1: Kompetenz-orientierung		Modell 2: Differen-zierung		Modell 3: Verengung des Lehrplans		Modell 4: Keine Ver-änderung	
	b	*SE*	*b*	*SE*	*b*	*SE*	*b*	*SE*
Lehrermerkmale								
Unterrichtsentwicklung	**0.52**	0.03	**0.38**	0.03	**0.13**	0.03	**−0.33**	0.03
Kontrolle	−0.02	0.03	0.01	0.03	**0.12**	0.03	0.05	0.03
Geschlecht[1]	−0.03	0.06	0.11	0.07	−0.13	0.07	**0.20**	0.07
Berufserfahrung[2]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Schülermerkmale (auf Klassenebene aggregiert)								
Familiensprache[3]	−0.02	0.02	−0.01	0.02	0.02	0.02	−0.03	0.02
sozio-ökonomischer Status (HISEI)[4]	−0.04	0.05	−0.03	0.05	−0.07	0.05	**0.14**	0.05
R^2	.28		.16		.04		.12	
N	1757		1757		1757		1757	

Anmerkung. Fett gedruckte Regressionskoeffizienten *(b)* sind statistisch signifikant von 0 verschieden ($p < .05$).

[1] Geschlecht der Lehrkraft: 1 = weiblich, 0 = männlich

[2] Berufserfahrung in Jahren

[3] Familiensprache auf Klassenebene aggregiert (Einheit in 10 % skaliert)

[4] sozio-ökonomischer Status erfasst mit dem HISEI auf Grundlage des ISCO 08, auf Klassenebene aggregiert.

Tab. 3: Regressionsanalyse zur Vorhersage selbstberichteter Unterrichtsveränderungen anhand von Überzeugungen von Lehrkräften zu den Funktionen von VERA und weiteren Kontrollvariablen (Regressionen ausschließlich auf Klassenebene)

Status der Schülerschaft innerhalb der Klasse steht in Beziehung zu ausbleibenden Unterrichtsmodifikationen ($b = 0.14$, $p < .05$).

In einem weiteren Schritt wurden die Überzeugungen der Lehrkräfte zur Vorhersage der Kompetenzen von Schülerinnen und Schülern im Lesen und in Mathematik verwendet (vgl. Tabelle 4). Pro Kompetenzbereich wurde ein Mehrebenenmodell geschätzt, in das neben den Überzeugungen der Lehrkräfte verschiedene Kovariaten der Schülerinnen und Schüler sowie der Lehrkräfte eingingen. Zu den Schülermerkmalen gehört neben der Familiensprache auch der sozio-ökonomische Status, wobei diese Angaben sowohl auf Individual- als auch auf Klassenebene im Modell berücksichtigt wurden. Als Hintergrundmerkmale der Lehrkräfte wurden – wie in der Regressionsanalyse – das Geschlecht und die Berufserfahrung kontrolliert. In Modell 1 besteht nach Kontrolle aller aufgeführten Hintergrundmerkmale ein schwach positiver, aber statistisch signifikanter Zusammenhang zwischen der Einschätzung von VERA als Unterrichtsentwicklungsinstrument und der von den Schülerinnen und Schülern erreichten Lesekompetenz ($b = 3.7$, $p < .05$). Ein Anstieg in der Skala *Unterrichtsentwicklung* um eine Standardabweichung geht demnach mit einer durchschnittlichen Erhöhung der Leseleistung von

Prädiktoren	Modell 1: Lesekompetenz		Modell 2: Mathematische Kompetenzen	
	b	*SE*	*b*	*SE*
Individualebene				
Familiensprache[1]	**−22.5**	3.3	**−26.3**	2.9
sozio-ökonomischer Status[2]	**28.7**	1.2	**26.0**	1.2
Klassenebene				
Lehrermerkmale				
Unterrichtsentwicklung	**3.7**	1.6	**9.3**	1.8
Kontrolle	0.6	1.8	2.9	1.7
Geschlecht[3]	2.9	4.5	4.4	3.6
Berufserfahrung[4]	0.1	0.1	0.2	0.1
Schülermerkmale				
Familiensprache[5]	**−13.4**	1.6	**−13.9**	1.4
sozio-ökonomischer Status[6]	**46.1**	4.7	**55.1**	5.0
R^2 (Individualebene)	.10		.09	
R^2 (Klassenebene)	.49		.57	
N	22 389		22 002	

Anmerkung. Fett gedruckte Regressionskoeffizienten *(b)* sind statistisch signifikant von 0 verschieden ($p < .05$).
[1] Familiensprache: 1 = manchmal Deutsch oder nie Deutsch, 0 = immer Deutsch
[2] sozio-ökonomischer Status erfasst mit dem HISEI auf Grundlage des ISCO 08 (zentriert am Gesamtwert)
[3] Geschlecht der Lehrkraft: 1 = weiblich, 0 = männlich
[4] Berufserfahrung in Jahren
[5] Familiensprache auf Klassenebene aggregiert (Einheit in 10 % skaliert)
[6] sozio-ökonomischer Status erfasst mit dem HISEI auf Grundlage des ISCO 08, auf Klassenebene aggregiert.

Tab. 4: Mehrebenenanalyse zur Vorhersage der Lesekompetenz und der mathematischen Kompetenzen der Schülerinnen und Schüler anhand von Überzeugungen ihrer Lehrkräfte zu den Funktionen von VERA und weiteren Kontrollvariablen auf Klassen- und Individualebene

4 Punkten einher. In Modell 2, das die mathematischen Kompetenzen vorhersagt, zeigt sich ebenfalls ein signifikant positiver Zusammenhang mit einer durchschnittlich erwartbaren Zunahme von 9 Punkten bei entsprechend höherer Wahrnehmung von VERA als Instrument der Unterrichtsentwicklung.

Erwartungsgemäß zeigen sich außerdem sowohl auf Individualebene als auch auf Klassenebene Zusammenhänge der als Kontrollvariablen berücksichtigten Familiensprache sowie des sozio-ökonomischen Status der Eltern: Innerhalb von Klassen schneiden Schülerinnen und Schüler, die zuhause nur Deutsch sprechen bzw. deren Eltern einen höheren sozio-ökonomischen Status aufweisen, besser in den Kompetenztests im Lesen und in Mathematik ab. Unter Kontrolle dieser Individualmerkmale zeigen sich auch systematische Zusammenhänge dieser Merkmale auf Klassenebene. In Klassen mit einem höheren Anteil von Familien, die zuhause nur Deutsch sprechen, erreichen Schülerinnen und Schüler bessere Ergebnisse im Lesen und in Mathematik. Darüber hinaus geht ein durchschnittlich höherer sozio-ökonomischer Status in Klassen mit besseren Leistungen in beiden Domänen einher.

4. Diskussion

Das Ziel dieser Arbeit war es, die aus Sicht von Lehrkräften wahrgenommenen Funktionen der Vergleichsarbeiten zu beschreiben und die Zusammenhänge dieser Wahrnehmungen mit der Unterrichtsgestaltung und den schülerseitig erreichten Kompetenzen zu untersuchen. Es wurde angenommen, dass Lehrkräfte, die VERA eher als ein Instrument zur Unterrichtsentwicklung betrachten, ihren Unterricht stärker im Sinne der Bildungsstandards weiterentwickeln und in ihren Klassen bessere Leistungsergebnisse erzielen. Weiterhin wurde vermutet, dass Lehrkräfte, die VERA als ein Kontrollinstrument wahrnehmen, in ihren Klassen zwar ebenfalls verstärkt auf gute Schülerleistungen hinwirken, sich aber weniger um einen kompetenzorientierten und differenzierten Unterricht bemühen.

Die Mittelwerte der beiden Skalen zur Erfassung der wahrgenommenen Funktionen befanden sich jeweils nahe beim theoretischen Mittelwert der Skala. Dies lässt darauf schließen, dass die Gesamtheit aller Lehrkräfte dem VERA-Test weder die eine noch die andere Funktion klar zuschreibt. Da jedoch die Funktion der Unterrichts- und Schulentwicklung vonseiten der Kultusministerkonferenz wiederholt hervorgehoben wurde (vgl. KMK, 2006, 2010, 2012), sollte sich eine stärkere Polarisierung in den Überzeugungen der Lehrkräfte zeigen. Es scheint daher angezeigt, die primäre Funktion von VERA deutlicher zu kommunizieren, um Lehrkräfte für das Potenzial des Tests und der Rückmeldungen zu sensibilisieren. Die deskriptiven Ergebnisse der Skalen zu den vorgenommenen Veränderungen im Unterricht deuten zunächst an, dass die Einführung von Leistungstests aus Sicht von Lehrkräften nicht wesentlich zu einer verstärkten Kompetenzorientierung und Differenzierung beigetragen hat, aber auch nicht zu einer substanziellen Verengung des Curriculums führte. Dieser Befund stimmt mit den Ergebnissen einer baden-württembergischen Studie überein, in der Realschullehrkräfte keine

Verengung des Curriculums infolge der Einführung von VERA berichteten (Wacker & Kramer, 2012).

Zentral für die vorliegende Arbeit ist der Zusammenhang zwischen den erfragten Überzeugungen der Lehrkräfte und der Unterrichtsgestaltung sowie den schülerseitigen Kompetenzen. Hier weisen die Ergebnisse darauf hin, dass Lehrkräfte, die die Vergleichsarbeiten als Mittel der Unterrichtsentwicklung ansehen, stärker dazu tendieren, ihren Unterricht auf den Erwerb von Kompetenzen auszurichten und auf die Unterschiede in der Klasse mit Differenzierungsmaßnahmen einzugehen. Wenn Lehrkräfte also VERA mit der Funktion assoziieren, die der Test primär erfüllen soll, berichten diese häufiger von Veränderungen in ihrem Unterricht, die den Zielen von VERA entsprechen. Dieser Befund stützt somit die vorab angenommene Hypothese. Die Wahrnehmung der Entwicklungs- bzw. Kontrollfunktionen von VERA hängt aber auch mit der Wahrnehmung zusammen, dass sich das im eigenen Unterricht implementierte Curriculum in seiner Breite reduziert hat. Stimmen Lehrkräfte einer der beiden untersuchten Funktionen zu, so geht dies in der Tendenz mit einer stärkeren Verengung des Curriculums einher. Zwar handelt es sich bei diesen Zusammenhängen um kleine Effekte (vgl. Cohen, 1969), die wenig erklärungsmächtig sind, jedoch sollte dies Gegenstand weiterer Forschung sein, inwiefern sich hier unerwünschte negative Effekte manifestieren.

Die vorliegende Studie konnte ebenfalls Belege dafür finden, dass die Überzeugungen zu den Vergleichsarbeiten mit den Leistungen der Schülerinnen und Schüler in einem bedeutsamen Zusammenhang stehen, der auch unter Berücksichtigung zentraler individueller und klassenbezogener Hintergrundmerkmale signifikant ist. Dieser positive Zusammenhang findet sich allerdings nur, wenn VERA von den betreffenden Lehrkräften als Instrument der Unterrichtsentwicklung verstanden wird. Zur Erklärung dieses Zusammenhangs lassen sich zwei Vermutungen aufstellen. Ein erster Erklärungsansatz bezieht sich auf den Umgang mit Daten aus Vergleichsarbeiten. Lehrkräfte, für die VERA ein Instrument der Unterrichtsentwicklung darstellt, nutzen die Rückmeldungen auch stärker zur Reflexion des eigenen Unterrichts als Lehrkräfte, für die das Instrument zur Kontrolle von Schulen dient (vgl. Kühle & Peek, 2007). Verwenden Lehrkräfte die Ergebnisse von VERA in Jahrgangsstufe 3 dazu, Defizite in den Leistungsständen zu diagnostizieren und den Unterricht im anschließenden Schuljahr dementsprechend anzupassen, so sollte sich dies in den Ergebnissen des IQB-Ländervergleichs zum Ende der Jahrgangsstufe 4 niederschlagen und zu den gefundenen Zusammenhängen führen. Ein zweiter Erklärungsansatz beruht auf dem Befund dieser Arbeit, dass die Wahrnehmung von VERA als Unterrichtsentwicklungsinstrument mit einer stärkeren Kompetenzorientierung im Unterricht einhergeht. Es könnte insofern vermutet werden, dass die Unterrichtsqualität bei Lehrkräften, die VERA als Mittel der Unterrichtsentwicklung begreifen, höher ist als bei Lehrkräften, für die VERA ein Kontrollinstrument ist. Es bedarf daher in zukünftigen Studien vermehrt Informationen über die Qualität des Unterrichts, um untersuchen zu können, in welcher Beziehung die Wahrnehmung von VERA und die Unterrichtsgestaltung durch die Lehrkraft stehen.

Der vermutete positive Zusammenhang zwischen der Wahrnehmung der Vergleichsarbeiten als Kontrollinstrument und höheren schülerseitigen Kompetenzständen ließ

sich nicht bestätigen. Dies könnte als Hinweis dafür interpretiert werden, dass Lehrkräfte, die davon überzeugt sind, dass Vergleichsarbeiten eine Kontrollfunktion erfüllen, nicht verstärkt auf die Kompetenzentwicklung der Schülerinnen und Schüler einwirken. Die Datenerhebung des IQB-Ländervergleichs, welche die Grundlage unserer Studie bildet, wurde von externen Testleitern unter standardisierten Bedingungen durchgeführt, sodass eine gezielte Vorbereitung auf den Test und eine Einflussnahme auf die Testergebnisse nicht möglich waren. Bei VERA sind die Lehrkräfte in der Regel selbst für die Vorbereitung, Durchführung und Auswertung verantwortlich und besitzen somit Spielräume, das Testergebnis ggf. positiv zu beeinflussen. Es ist daher in weiteren Studien zu prüfen, welcher Zusammenhang zwischen Kontrollüberzeugungen von Lehrkräften und den Ergebnissen in VERA selbst besteht und ob er sich vom hier berichteten unterscheidet.

Eine Einschränkung unserer Studie besteht darin, dass es sich bei dieser Erhebung um eine Querschnittserhebung handelt, die korrelative und keine kausalen Beziehungen beschreibt. Es ist deshalb mit den vorliegenden Daten nicht möglich festzustellen, ob die Wahrnehmung der Entwicklungsfunktion von VERA ursächlich den beobachteten Leistungsvorteil bedingt. Die relative Stabilität von Überzeugungen (vgl. Pajares, 1992) spricht jedoch dafür, dass die hier erfassten Merkmale, vermittelt über Drittvariablen, die Leistungen bedingen und nicht umgekehrt. Ferner wurden in dieser Studie ausschließlich Grundschullehrkräfte befragt, sodass offen bleibt, inwiefern sich die Ergebnisse auch auf die Schularten der Sekundarstufe I übertragen lassen, in denen VERA in der 8. Jahrgangsstufe durchgeführt wird. Eine weitere Einschränkung besteht darin, dass die Veränderungen im Unterricht nicht fachspezifisch erhoben wurden und somit keine fachspezifischen Zusammenhänge zu den Überzeugungen untersucht werden konnten. Darüber hinaus beschränken sich die Ergebnisse der vorliegenden Studie auf ausgewählte Domänen des IQB-Ländervergleichs in der Grundschule. Inwiefern unsere Befunde auch auf die Ergebnisse anderer Tests (z. B. bei VERA selbst) oder andere Kompetenzbereiche übertragbar sind, lässt sich nicht beantworten.

Abschließend kann festgehalten werden, dass nicht nur die häufig untersuchten Globalbeurteilungen von VERA (z. B. Akzeptanz und Nützlichkeit) von Bedeutung zu sein scheinen, sondern auch die Überzeugungen der Lehrkräfte hinsichtlich der Funktionen dieses Instruments. Damit die Vergleichsarbeiten ihre intendierte positive Funktion als Element der Schul- und Unterrichtsentwicklung entfalten können, sollten nach wie vor Anstrengungen unternommen werden, diese Funktion von VERA allen schulischen Akteuren transparent zu machen.

Literatur

Altrichter, H., & Maag-Merki, K. (2010). Steuerung der Entwicklung des Schulwesens. In H. Altrichter & K. Maag-Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 16–39). Wiesbaden: VS Verlag für Sozialwissenschaften.

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.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O., & Neubrand, J. (1997). *TIMSS – Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich: Deskriptive Befunde.* Opladen: Leske + Budrich.

Bellmann, J., & Weiß, M. (2009). Risiken und Nebenwirkungen Neuer Steuerung im Schulsystem. *Zeitschrift für Pädagogik, 55*(2), 286–308.

Böhme, K., & Bremerich-Vos, A. (2012). Beschreibung der im Fach Deutsch untersuchten Kompetenzen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 19–33). Münster: Waxmann.

Bonsen, M., Büchter, A., & Peek, R. (2006). Datengestützte Schul- und Unterrichtsentwicklung. Bewertungen der Lernstandserhebungen in NRW durch Lehrerinnen und Lehrer. *Jahrbuch der Schulentwicklung, 14*, 125–148.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*(1), 67–90.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Dedering, K. (2011). Hat Feedback eine positive Wirkung? Zur Verarbeitung extern erhobener Leistungsdaten in Schulen. *Unterrichtswissenschaft, 39*(1), 63–83.

Ganzeboom, H. B. G., de Graaf, P. M., Treiman, D. J., & de Leeuw, J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21*, 1–56.

Griffith, G., & Scharmann, L. (2008). Initial impacts of No Child Left Behind on elementary science education. *Journal of Elementary Science Education, 20*(3), 35–48.

Groß Ophoff, J., Koch, U., Hosenfeld, I., & Helmke, A. (2006). Ergebnisrückmeldungen und ihre Rezeption im Projekt VERA. In H. Kuper (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungsbereich* (S. 19–40). Münster: Waxmann.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J., Naftel, S., & Barney, H. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states* (No. Paperback ISBN/EAN: 978-0-8330-4149-4). Santa Monica, CA: RAND Corporation.

Helmke, A. (2004). Von der Evaluation zur Innovation: Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. *Das Seminar, 2*, 90–112.

Helmke, A., & Hosenfeld, I. (2003). Vergleichsarbeiten (VERA): Eine Standortbestimmung zur Sicherung schulischer Kompetenzen – Teil 1: Grundlagen, Ziele, Realisierung. *SchulVerwaltung, Ausgabe Hessen/Rheinland-Pfalz/Saarland, 1*, 10–13.

KMK (2005) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (2005). *Bildungsstandards der Kultusministerkonferenz: Erläuterungen zur Konzeption und Entwicklung*. Neuwied: Luchterhand.

KMK (2006) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. München: Luchterhand.

KMK (2010) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2010). *Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung*. Köln: Wolters Kluwer.

KMK (2012) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2012). *Vereinbarung zur Weiterentwicklung von VERA*. Berlin: KMK.

KMK (2013) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2013). *VERA 3 und VERA 8 (Vergleichsarbeiten in den Jahrgangsstufen 3 und 8): Fragen und Antworten für Schulen und Lehrkräfte*. Berlin: KMK.

Kober, N., Chudowsky, N., & Chudowsky, V. (2008). *Has student achievement increased since 2002? State test score trends through 2006–07*. Washington, DC: Center on Educational Policy.

Koch, U., Groß Ophoff, J., Hosenfeld, I., & Helmke, A. (2006). Qualitätssicherung: Von der Evaluation zur Schul- und Unterrichtsentwicklung – Ergebnisse der Lehrerbefragungen zur Auseinandersetzung mit den VERA-Rückmeldungen. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 187–199). Münster: Waxmann.

Kühle, B., & Peek, R. (2007). Lernstandserhebungen in Nordrhein-Westfalen: Evaluationsbefunde zur Rezeption und zum Umgang mit Ergebnisrückmeldungen in Schulen. *Empirische Pädagogik, 21*(4), 428–447.

Kuper, H., & Hartung, V. (2007). Überzeugungen zur Verwendung des Wissens aus Lernstandserhebungen. Eine professionstheoretische Analyse. *Zeitschrift für Erziehungswissenschaft, 2*(10), 214–229.

Lee, J. (2007). Do national and state assessments converge for educational accountability? A meta-analytic synthesis of multiple measures in Maine and Kentucky. *Applied Measurement in Education, 20*, 171–203.

Maag-Merki, K. (2010). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter & K. Maag-Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 145–169). Wiesbaden: VS Verlag für Sozialwissenschaften.

Maier, U. (2007). Welche Konsequenzen ziehen Mathematiklehrkräfte aus verpflichtenden Diagnose- und Vergleichsarbeiten? *mathematica didactica, 30*(2), 5–31.

Maier, U., & Kuper, H. (2012). Vergleichsarbeiten als Instrumente der Qualitätsentwicklung an Schulen. *Die deutsche Schule, 104*(1), 88–99.

Maier, U., Metz, K., Bohl, T., Kleinknecht, M., & Schymala, M. (2012). Vergleichsarbeiten als Instrument der datenbasierten Schul- und Unterrichtsentwicklung in Gymnasien. Empirische Befunde und forschungsmethodische Implikationen. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung* (S. 197–224). Wiesbaden: VS Verlag für Sozialwissenschaften.

McMurrer, J. (2007). *Choices, changes, and challenges: Curriculum and instruction in the NCLB era*. Washington, DC: Center on Education Policy.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus statistical analysis with latent variables: User's guide*. Los Angeles: Muthén & Muthén.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*(3), 307–332.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2. Aufl.). Thousand Oaks, CA: Sage Publications.

Rentner, D. S., Scott, C., Kober, N., Chudowsky, N., Chudowsky, V., Joftus, S., & Zabala, D. (2006). *From the capital to the classroom: Year 4 of the No Child Left Behind Act*. Washington, DC: Center on Education Policy.

Richter, D., Engelbert, M., Böhme, K., Haag, N., Hannighofer, J., Reimers, H., Roppelt, A., Weirich, S., Pant, H. A., & Stanat, P. (2012). Anlage und Durchführung des Ländervergleichs. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 85–102). Münster: Waxmann.

Roppelt, A., & Reiss, K. (2012). Beschreibung der im Fach Mathematik untersuchten Kompetenzen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 34–43). Münster: Waxmann.

Schneewind, J., & Kuper, H. (2009). Rückmeldeformate und Verwendungsmöglichkeiten der Ergebnisse aus zentralen Lernstandserhebungen. In T. Bohl & H. Kiper (Hrsg.), *Lernen aus Evaluationsergebnissen* (S. 113–129). Bad Heilbrunn: Klinkhardt.

Smith, J. M., & Kovacs, P. E. (2011). The impact of standards-based reform on teachers: the case of ‚No Child Left Behind'. *Teachers and Teaching: Theory and Practice, 17*(2), 201–225.

Stanat, P., Pant, H. A., Böhme, K., & Richter, D. (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series, 4*, 9–36.

Wacker, A., & Kramer, J. (2012). Vergleichsarbeiten in Baden-Württemberg. *Zeitschrift für Erziehungswissenschaft, 15*(4), 683–706.

Weirich, S., Haag, N., & Roppelt, A. (2012). Testdesign und Auswertung des Ländervergleichs: Technische Grundlagen. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik* (S. 277–290). Münster: Waxmann.

Abstract: Comparative tests based on the educational standards set by the Conference of Education Ministers (*Vergleichsarbeiten;* German abbreviation: VERA) have for many years been an important instrument in competence diagnostics. They serve primarily school and curriculum development; sometimes, however, they are also used for the comprehensive information of the school supervision authorities concerning the performance level of individual schools. The present contribution examines how far teachers are aware of these functions and in what way they are related to characteristics of classroom instruction and to the students' competencies. The study is based on data provided by the national assessment study carried out by the IQB (Institute for Educational Quality Improvement) in 2011 on the level of primary education, which focused on competencies in the subjects math and German. The analyses show that teachers who consider VERA a means for the development of teaching orientate their lessons much more towards the development of competencies and undertake a more detailed differentiation in their lessons. Furthermore, students of such teachers achieve better results in both reading and math, even after allowance for individual and class-related background characteristics.

Keywords: Functions of Student Achievement Tests, Elementary School, National Assessment Study, Reading Literacy, Mathematical Literacy

Anschrift der Autor(inn)en

Dr. Dirk Richter, Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen, Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: dirk.richter@iqb.hu-berlin.de

Dr. Katrin Böhme, Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen, Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: katrin.boehme@iqb.hu-berlin.de

Dr. Michael Becker, Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Abteilung Steuerung und Finanzierung des Bildungswesens, Schloßstraße 29, 60486 Frankfurt am Main, Deutschland
E-Mail: becker@dipf.de

Prof. Dr. Hans Anand Pant, Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen, Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: iqboffice@iqb.hu-berlin.de

Prof. Dr. Petra Stanat, Humboldt-Universität zu Berlin, Institut zur Qualitätsentwicklung im Bildungswesen, Unter den Linden 6, 10099 Berlin, Deutschland
E-Mail: iqboffice@iqb.hu-berlin.de