Richter, Dirk; Böhme, Kathrin; Becker, Michael; Pant, Hans Anand; Stanat, Petra

Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten. Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen von Schülerinnen und Schülern

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Zeitschrift für Pädagogik 60 (2014) 2, S. 225-244

Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten: Zusammenhänge zu Veränderungen im Unterricht und den Kompetenzen der Schülerinnen und Schüler

Dirk Richter[1], Katrin Böhme[1], Michael Becker[2], Hans Anand Pant[1], Petra Stanat[1]

[1] Institut zur Qualitätsentwicklung im Bildungswesen, Humboldt-Universität zu Berlin
[2] Universität Potsdam

Zusammenfassung

Die Vergleichsarbeiten (VERA) bilden 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 der 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 Hintergrundvariablen.

1. Einleitung

Nach dem erwartungswidrig schlechten Abschneiden Deutschlands in den Schulleistungsstudien TIMSS (Baumert et al., 1997) und PISA (Baumert et al., 2001) setzte ein umfassender Reformprozess im Bildungswesen ein, der darauf abzielte, das input-orientierte Steuerungsmodell durch Elemente einer output-orientierten Steuerung zu ergänzen (z.B. Altrichter & Maag-Merki, 2010). Eine wesentliche Grundlage dieser Optimierungsbemühungen der Kultusministerkonferenz (KMK) bildet die Einführung länderübergreifend verbindlicher Bildungsstandards, die beschreiben, welche Kompetenzen Schülerinnen und Schüler bis zum Ende eines Bildungsabschnitts erworben haben sollen (Klieme et al., 2003). Eine Möglichkeit der 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 Durchführung landesweiter 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 kann sich auf die Unterstützung der Arbeit der Schulaufsicht und/oder der Schulinspektion beziehen, die in einigen Ländern Einsicht in die VERA-Ergebnisse auf Schul- und Klassenebene erhalten (KMK, 2012). Somit nimmt VERA – je nach Landesregelung – eine Doppelfunktion ein: Die Tests dienen einerseits der Schul- und Unterrichtsentwicklung, vermittelt über eine Evaluation von Schülerleistungen auf der Ebene der Klasse oder Schule, und andererseits der Rechenschaftslegung in Bezug auf Schülerleistungen auf der Ebene einzelner Klassen, Schulen, für Schulen bestimmter Schularten, für Regionen oder für ein Landes insgesamt.

[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.

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 bereits in den Anfangsjahren der Vergleichsarbeiten entstanden ist, ist unklar, ob Lehrkräfte gut 10 Jahre nach der Einführung von VERA nach wie vor ähnliche Überzeugungen aufweisen. Insbesondere lässt sich aus den bisher vorliegenden Arbeiten nicht ableiten, ob die Überzeugungen der Lehrkräfte relevant für den Umgang mit Leistungsrückmeldungen und die Unterrichtsgestaltung sind. Einige Befunde weisen darauf hin, dass die Einschätzung der Nützlichkeit und der Akzeptanz von VERA für die Ableitung von unterrichtsbezogenen Maßnahmen eine Rolle spielt (Kühle & Peek, 2007; Maier, 2008) und somit individuelle Überzeugungen zu den Vergleichsarbeiten potenziell bedeutsam für das professionelle Handeln von Lehrkräften sein können. Es ist jedoch ungeklärt, welche Funktionen Lehrkräfte den Vergleichsarbeiten zuschreiben und wie sich diese Wahrnehmungen im unterrichtlichen Verhalten und in den Leistungen der Schülerinnen und Schüler niederschlagen.

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 auf 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.

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

Die Vergleichsarbeiten (VERA) im Primarbereich wurden erstmals 2003 in Rheinland-Pfalz im Fach Mathematik durchgeführt. Die Koordinierung lag damals bei der Projektgruppe *Empirische Bildungsforschung* an der Universität Koblenz-Landau (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 ergänzende Informationen hinsichtlich des Leistungsstands ihrer Schülerinnen und Schüler und können ggf. zusätzliche Förderangebote bereitstellen. Inzwischen – 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 jeweiligen VERA-Tests einer repräsentativen Stichprobe von Schülerinnen und Schülern gemeinsam mit bereits normierten Aufgaben zu den Bildungsstandards vorgelegt. Auf der Basis dieser gemeinsamen Datenerhebung gelingt eine psychometrische Anbindung von VERA an die Metrik der Bildungsstandards. Hierdurch lassen sich die empirischen Schwierigkeiten der VERA-Aufgaben und somit auch die Leistungen der Schülerinnen und Schüler auf verschiedenen Niveaus sogenannter Kompetenzstufenmodelle abbilden. Diese beschreiben, welche Kompetenzen eine Schülerin bzw. ein Schüler bereits erreicht hat. 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 angelegt, d.h. die Untersuchung des Kompetenzstandes erfolgt in allen Schulen und Klassen innerhalb eines Landes (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 hinaus gehende Veröffentlichungen der Ergebnisse von Einzelschulen, etwa in Form von Rankings, finden nicht statt (KMK, 2012).

Somit entspricht VERA einem Evaluationsverfahren, dass sich am Modell der professionellen Qualitätsentwicklung und Qualitätssicherung orientiert (Baumert, 2001) und dazu dienen soll, zentrale Aspekte der Bildungsstandards zu transportieren sowie Prozesse der Schul- und Unterrichtsentwicklung anzustoßen.

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 Funktionen VERA aus Sicht von Lehrkräften erfüllt, wurde bislang in nur wenigen Arbeiten empirisch untersucht. In einer qualitativen Studie mit Lehrkräften 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. Ein deskriptiver Vergleich der erfassten Skalen machte deutlich, 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; KMK, 2013) könnte dazu beigetragen haben, dass auch Lehrkräfte den Test 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 Testergebnissen sowohl zu einer Verbesserung der Unterrichtsqualität als auch der Lernergebnisse beiträgt (KMK, 2010, 2012). Rezeptionsstudien, die die Auseinandersetzung mit VERA-Rückmeldungen untersuchten, haben gezeigt, dass nur ein geringer Anteil von Lehrkräften die Rückmeldungen zur Ableitung von Maßnahmen für den eigenen Unterricht heranzieht (Bonsen, Büchter & Peek, 2006; Dedering, 2011; Groß Ophoff, Koch, Hosenfeld

& Helmke, 2006; Kühle & Peek, 2007; Maier, 2007). Geben Lehrkräfte an, dass sie ihren Unterricht auf Grundlage der VERA-Rückmeldungen modifizieren, so beziehen sich diese Veränderungen vor allem auf die Verwendung von Aufgaben aus dem Test, eine veränderte Unterrichtsgestaltung (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 internationalen Studien ist bekannt, dass auch High-Stakes-Tests, wie sie in den USA durchgeführt werden, zu positiven Konsequenzen in der Unterrichtsgestaltung beitragen können (Maag-Merki, 2010). Befragungen von Lehrkräften weisen darauf hin, dass sie nach der Einführung von High-Stakes-Tests mehr Zeit für das Unterrichten von Fachinhalten aufwenden, nach effektiveren Unterrichtsmethoden suchen, Aufgaben differenziert nach Fähigkeitsniveau auswählen und schwierige Themen in der Klasse wiederholen (Hamilton et al., 2007). Lehrkräfte scheinen somit durchaus die Rückmeldungen über die Leistungen ihrer Schülerinnen und Schüler für eine optimierte Gestaltung von Lernprozessen zu nutzen.

In der öffentlichen Diskussion über VERA wird gelegentlich die Befürchtung geäußert, dass mit den Tests auch negative Folgen verbunden sein könnten. Diese Befürchtung bezieht sich beispielsweise auf das Üben von Testaufgaben, die Einbeziehung von Testaufgaben in den Unterricht und die Hilfestellung bei der Bearbeitung der Tests (Brügelmann, 2005; Stähling, 2005). Bislang gibt es nur wenige Studien, die dieser Vermutung in Deutschland empirisch nachgegangen sind, Befunde einer baden-württembergischen Lehrkräftebefragung legen jedoch nahe, dass die Vergleichsarbeiten zwar nicht zu einer starken Verengung des Curriculums beitragen, gleichzeitig aber auch kaum zur Unterrichtsplanung und Diagnose von Lernrückständen genutzt werden (Wacker & Kramer, 2012).

Weitaus umfangreichere Erkenntnisse über die negativen Folgen von Schulleistungstests liegen jedoch aus US-amerikanischen Studien vor, in denen die Auswirkungen des *No-Child-Left-Behind-Gesetzes* (NCLB) auf den Unterricht analysiert wurden (Bellmann & Weiß, 2009; Hamilton et al., 2007; Maag-Merki, 2010; Maier & Kuper, 2012). Übereinstimmend zeigen mehrere Studien, dass sich die Unterrichtszeit in geprüften Fächern (Englisch und Mathematik) deutlich erhöhte, während sich diese in anderen Fächern reduzierte (Griffith & Scharmann, 2008; Hamilton et al., 2007; McMurrer, 2007; Rentner et al., 2006; Smith & Kovacs, 2011). Darüber hinaus liegen Hinweise auf 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). Neben der inhaltlichen

Angleichung von Unterricht und Test kam es auch zu Angleichungen bei den verwendeten Prüfungsformaten im Unterricht, um Schülerinnen und Schüler mit der Art der Aufgaben vertraut zu machen, die in den Tests zum Einsatz kommen (Hamilton et al., 2007).

Die verstärkte Ausrichtung des Unterrichts auf die Inhalte der High-Stakes-Tests scheint sich auch in den Schülerleistungen widerzuspiegeln. Bei den High-Stakes-Tests zeigte sich in der Mehrzahl der Bundesstaaten nach der Einführung von *No Child Left Behind* ein Anstieg der als *proficient* eingeschätzten Schülerinnen und Schüler (Kober, Chudowsky & Chudowsky, 2008). Im stichprobenbasierten *National Assessment of Educational Progress (NAEP)* konnten jedoch in vergleichbaren Domänen (Englisch.-Lesen und Mathematik) meist nur geringere Leistungssteigerungen verzeichnet werden (Jacob, 2007; Kober et al., 2008; Lee, 2007). Dieser differenzielle Leistungsanstieg wird auch als *inflated gain score* bezeichnet und deutet darauf hin, dass die Zuwächse im High-Stakes-Test nur eingeschränkt eine tatsächliche, über verschiedene Tests generalisierbare Kompetenzsteigerung abbilden.

Die in den USA beobachteten Auswirkungen von High-Stakes-Test lassen sich nicht in gleichem Maße für den deutschen VERA-Test antizipieren, weil das Ergebnis dieses Tests keine direkten Konsequenzen für Schülerinnen und Schüler, Lehrkräfte und Schulen mit sich bringt. Da die Vergleichsarbeiten neben der Unterrichts- und Schulentwicklung in einigen Ländern aber auch zur Rechenschaftslegung von Schülerleistungen auf Klassen- und Schulebene verwendet werden, lässt sich vermuten, dass sich Lehrkräfte, die VERA als Instrument der Rechenschaftslegung wahrnehmen, bemühen, mit ihren Klassen im Test gut abzuschneiden. Inwiefern die wahrgenommene Funktion der Vergleichsarbeiten mit dem unterrichtlichen Verhalten der Lehrkraft und den Leistungen der Schülerinnen und Schüler zusammenhängt, ist bislang noch nicht bekannt. Es gibt jedoch Hinweise aus einer Untersuchung in Nordrhein-Westfalen, dass Lehrkräfte, die Vergleichsarbeiten als Mittel zur Schul- und Unterrichtsentwicklung ansehen, auch besser über den Test informiert sind und ihn als bedeutsamer, nützlicher und aufschlussreicher einschätzen (Kühle & Peek, 2007).

1.3 Fragestellung der Studie

Die vorliegende Arbeit untersucht die Wahrnehmung der durch VERA übernommenen 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 kann angenommen werden, dass Lehrkräfte, die VERA als ein Mittel zur Unterrichtsentwicklung begreifen, die Ergebnisse des Tests verstärkt zur Reflexion des Unterrichts nutzen und den Erwerb von Kompetenzen gezielt fördern. Lehrkräfte, die VERA vor allem als Kontrollinstrument betrachten, sollten ebenfalls Veränderungen zeigen, die eine Verbesserung der Testergebnisse erwarten lassen. Spezifische Hypothesen über das zu erwartende Verhalten lassen sich jedoch nicht aus der Literatur ableiten. 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 im Primarbereich des Jahres 2011, der anhand einer national repräsentativen Stichprobe Kompetenzen von Schülerinnen und Schülern der vierten Jahrgangsstufe überprüfte (Stanat et al., 2012). Eingesetzt wurden Tests im Fach Deutsch zu den Bereichen Lesen und Zuhören sowie in Mathematik. Weiterhin beinhaltete die Studie eine Befragung aller Lehrkräfte der beteiligten Klassen in den Fächern Deutsch und Mathematik. Insgesamt nahmen an der Untersuchung jeweils eine 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 22389 von 26029 Schülerinnen und Schülern Fragebögen von Lehrkräften vor. Dies entspricht einem Anteil von 86.0 Prozent. Diese Kinder waren zum Zeitpunkt der Erhebung im Durchschnitt 10.5 Jahre alt (*SD* = 0.5 Jahre) und 49.6 Prozent von ihnen sind weiblich. Im Fach Mathematik konnten den teilnehmenden Lehrkräften insgesamt 22002 von 26016 Schülerinnen und Schülern zugeordnet werden, was einer Quote von 84.6 Prozent entspricht. Diese Kinder waren ebenfalls im Durchschnitt 10.5 Jahre alt (*SD* = 0.5 Jahre) und 49.4 Prozent 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. Eine Arbeit von Pant und Richter (eingereicht) konnte zeigen, dass die Skalen eine zweidimensionale Struktur repräsentieren und beide Dimensionen nur schwach positiv miteinander korrelieren (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 wurde auf Grundlage einer Studie von Hamilton et al. (2007) entwickelt und 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 beiden erstgenannten Konstrukte *Kompetenzorientierung* und *Differenzierung* positiv zu bewertende Folgen von VERA, die beiden letztgenannten Skalen, *Verengung des Lehrplans* und *keine Veränderung*, hingegen ambivalente bzw. negative Folgen.

Tabelle 1 bitte etwa hier einfügen.

Leistungstests (Schülerebene). Weiterhin werden Ergebnisse der Kompetenztests des Ländervergleichs 2011 im Lesen und in Mathematik in die Analyse einbezogen. Zur Erfassung der Lesekompetenz beantworteten Schülerinnen und Schüler Fragen, die sich auf Sachtexte und literarische Texte bezogen (Böhme & Bremerich-Vos, 2012). Innerhalb von Regelschulen betrug die Testzeit für das Fach Deutsch insgesamt 80 Minuten, wobei sie für den Leseteil in Abhängigkeit vom Testheft zwischen 20 und 40 Minuten variierte. 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 4. 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. Die Aufgaben des Mathematiktests beinhalteten jeweils einen kurzen Stimulustext, gefolgt von mehreren Fragen. 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.

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 unter Verwendung von Mehrebenenmodellen (Raudenbush & Bryk, 2002). In Mehrebenenmodellen 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 berücksichtigt. 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 werden 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 werden soll, ob die Mittelwerte der Skalen 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 aufweisen. Für die Skala *keine Veränderung* ist ein vergleichsweise hoher Mittelwert zu beobachten, 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 die Schülerinnen und Schüler eingehen, denen eine Lehrkraft zugeordnet werden konnte und die an Regelschulen unterrichtet wurden. In dieser Teilstichprobe fallen die Mittelwerte etwas höher und die Standardabweichungen etwas geringer aus als in der Gesamtgruppe.

Tabelle 2 bitte etwa hier einfügen.

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 dienten das Geschlecht und die Berufserfahrung der Lehrkräfte sowie auf Klassenebene aggregierte Angaben zur Familiensprache und dem 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 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. 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 zu den Überzeugungen der Lehrkräfte berücksichtigten Kontrollvariablen gering. Lediglich im vierten Modell, das das Ausbleiben von Veränderungen in der Unterrichtsgestaltung untersucht, hat auch das Geschlecht der Lehrkraft einen Einfluss und zwar in dem Sinne, dass Frauen stärker als Männer dazu neigen, in ihrem Unterricht keine Veränderungen vorzunehmen (b=0.20, p<.05). Auch ein höherer sozio-ökonomischer Status der Schülerschaft innerhalb der Klasse steht in Beziehung zu ausbleibenden Unterrichtsmodifikationen (b=0.14, p<.05).

Tabelle 3 bitte etwa hier einfügen.

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 4 Punkten einher. In Modell 2, das die mathematische Kompetenz 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 nur Deutsch zu Hause 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 nur Deutsch zu Hause 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.

Tabelle 4 bitte etwa hier einfügen.

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 verstärkt auf gute Schülerleistungen hinwirken.

Die Mittelwerte der beiden Skalen zur Erfassung der wahrgenommenen Funktionen befanden sich jeweils nahe beim theoretischen Mittelwert der Skala. Dies lässt darauf schließen, dass Lehrkräfte weder die eine noch die andere der beiden Funktionen den VERA-Tests klar zuschreiben. Da jedoch die Funktion der Unterrichts- und Schulentwicklung vonseiten der Kultusministerkonferenz wiederholt hervorgehoben wurde (vgl. KMK, 2006, 2012; KMK, 2013), hätte man hier eine stärkere Polarisierung in den Überzeugungen der Lehrkräfte erwarten können. Die deskriptiven Ergebnisse der Skalen zu den vorgenommenen Veränderungen im Unterricht deuten an, dass die Einführung von Leistungstests aus Sicht von Lehrkräften zwar nicht zu einer verstärkten Kompetenzorientierung und Differenzierung beigetragen hat, aber auch nicht zu einer verstärkten Verengung des Curriculums. 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 zur Unterrichtsgestaltung und den schülerseitigen Kompetenzen. Hier zeigen die Ergebnisse, 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. Insofern scheint es angezeigt, die primäre Funktion von VERA deutlicher zu kommunizieren, um Lehrkräfte für das Potenzial des Tests und der Rückmeldungen zu sensibilisieren. Die Wahrnehmung der Entwicklungs- bzw. Kontrollfunktionen von VERA hängt in ähnlichem Maße mit der Wahrnehmung zusammen, dass sich das im eigenen Unterricht implementierte Curriculum in seiner Breite reduziert hat. Dies weist darauf hin, dass unabhängig davon, welche Funktionen mit VERA assoziiert werden, eine Konzentration der Unterrichtsinhalte aus Sicht der Lehrkräfte stattgefunden hat. Bei diesen Zusammenhängen handelt es sich jedoch um kleine Effekte (vgl. Cohen, 1969), die kaum erklärungsmächtig sind.

Die vorliegende Studie konnte ebenfalls Belege dafür finden, dass die Überzeugungen zu den Vergleichsarbeiten mit den Leistungen der Schülerinnen und Schülern in einem bedeutsamen Zusammenhang stehen, der auch unter Berücksichtigung zentraler individueller und klassenbezogener Hintergrundmerkmale signifikant ist. 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 kann deshalb 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 kann als Hinweis darauf interpretiert werden, dass Lehrkräfte, die davon überzeugt sind, dass Vergleichsarbeiten eine Kontrollfunktion erfüllen, nicht verstärkt auf die Kompetenzentwicklung der Schülerinnen und Schüler einwirken. Da der Leistungstest dieser Studie im Rahmen des IQB-Ländervergleichs 2011 durchgeführt wurde, bleibt die Frage offen, ob die lehrerseitigen Kontrollüberzeugungen im Zusammenhang mit den Ergebnissen im VERA-Test stehen. Die Datenerhebung des IQB-Ländervergleichs wurde von externen Testleitern unter hochstandardisierten Bedingungen durchgeführt, sodass eine gezielte Vorbereitung auf den Ländervergleich nicht möglich war. Bei VERA sind die Lehrkräfte i.d.R. selbst für die Vorbereitung, Durchführung und Auswertung verantwortlich und besitzen somit Spielräume das Testergebnis positiv zu beeinflussen. Es ist daher in weiteren Studien zu prüfen, welcher Zusammenhang zwischen Kontrollüberzeugungen von Lehrkräften und Ergebnissen in VERA besteht und ob er sich vom hier berichteten unterscheidet. Im Falle von Differenzen könnte dies auf unerwünschte oder negative Veränderungen und Reaktionen in Bezug auf VERA hinweisen.

Eine Einschränkung der Studie besteht darin, dass es sich bei dieser Erhebung um eine Querschnittserhebung handelt, die ausschließlich 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. Auch beschränken sich die Ergebnisse der Leistungstests auf ausgewählte Domänen des IQB-Ländervergleichs in der Grundschule. Inwiefern die Befunde dieser Arbeit auch auf die Ergebnisse anderer Tests (z.B. bei VERA) oder andere Kompetenzbereiche übertragbar sind, lässt sich mit dieser Studie nicht beantworten.

Abschließend kann festgehalten werden, dass die Studie anhand einer für Deutschland repräsentativen Stichprobe starke Zusammenhänge zwischen den wahrgenommenen Funktionen von VERA und der von Lehrkräften berichteten Gestaltung des Unterrichts feststellen konnte und eher kleine Zusammenhänge zu den Leistungen von Schülerinnen und Schülern. Es scheinen also nicht nur die häufig untersuchten Globalbeurteilungen von VERA (z.B. Akzeptanz und Nützlichkeit) von Bedeutung zu sein, sondern auch die Überzeugungen der Lehrkräfte über die Funktionen des Tests. Damit die Vergleichsarbeiten ihre intendierte positive Funktion als Element der Schul- und Unterrichtsentwicklung entfalten können, ist es von Bedeutung, dass auch die Lehrkräfte von dieser Funktion überzeugt sind. Daher sollten vermehrt 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. (2001). Vergleichende Leistungsmessung im Bildungsbereich. In *Zukunftsfragen der Bildung. Zeitschrift für Pädagogik. 43. Beiheft* (S. 13–36). Weinheim: Beltz.

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 und 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.

Brügelmann, H. (2005). Wahrheit durch Vera? Anmerkungen zum ersten Durchgang der landesweiten Leistungstests in sieben Bundesländern. *Grundschule aktuell, 89*, 7–9.

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.

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.

Jacob, B. (2007). *Test-based accountability and student achievement: an investigation of differential performance on NAEP and state assessments.* (No. Working Paper 12817). Cambrigde, MA.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J. & Tenorth, H.-E. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. Berlin: Bundesministerium für Bildung und Forschung.

KMK. (2006). Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring. München: LinkLuchterhand.

KMK. (2010). Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung. Köln: Wolters Kluwer.

KMK. (2012). Vereinbarung zur Weiterentwicklung von VERA. Berlin: KMK.

KMK. (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: Verlag für Sozialwissenschaften.

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

Maier, U. (2008). Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. *Zeitschrift für Pädagogik, 54*(1), 95–117.

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). *M plus statistical ananysis 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.

Pant, H. A. & Richter, D. (eingereicht). Funktionen von Vergleichsarbeiten und ihre Wahrnehmung aus der Perspektive von Lehrkräften.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2. ed Auflage). 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*.

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.

Stähling, R. (2005). Qualitätsentwicklung statt Vergleichsarbeiten. Zu einem unfruchtbaren Verhältnis von Forschung und Schule. *Die Deutsche Schule, 97*(2), 211–221.

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.

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

Skala	Itemzahl	Reliabilität Cronbachs Alpha	Beispielitem
Überzeugungen zu den Funktionen von Leistungstests			
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.

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

Variablen	*N*	*M*	*SD*	Schiefe	*t*-Test: μ=2.5 *p*-Wert
Überzeugungen zu den Funktionen von VERA					
Unterrichtsentwicklung	1691	2.42	0.64	-0.30	<.01
Kontrolle	1659	2.44	0.73	-0.19	<.01
Veränderungen im Unterricht					
Kompetenzorientierung	1657	2.42	0.68	-0.53	<.01
Differenzierung	1649	2.20	0.76	0.02	<.01
Verengung des Lehrplans	1662	1.70	0.57	0.52	<.01
Keine Veränderung	1574	2.71	0.76	-0.25	<.01
Kompetenztests					
Lesekompetenz	22389	504.87	95.89	-0.17	-
Mathematische Kompetenz	22002	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.

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

Prädiktoren	Modell 1: Kompetenz-orientierung		Modell 2: Differen-zierung		Modell 3: Verengung des Lehrplans		Modell 4: Keine Veränderung	
	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	0.28		0.16		0.04		0.12	
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.

Tabelle 4: Mehrebenenanalyse zur Vorhersage der Lese- und Mathematikkompetenz der Schülerinnen und Schüler anhand von Überzeugungen ihrer Lehrkräfte zu den Funktionen von VERA und weiteren Kontrollvariablen auf Klassen- und Individualebene

Prädiktoren	Modell 1: Lesekompetenz		Modell 2: Mathematik-kompetenz	
	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)	0.10		0.09	
R^2 (Klassenebene)	0.49		0.57	
N	22389		22002	

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.