Z Erziehungswiss (2012) 15:683–706
DOI 10.1007/s11618-012-0326-4

SCHWERPUNKT

ZfE

Vergleichsarbeiten in Baden-Württemberg

Zur Einschätzung von Lehrkräften vor und nach der Implementation

Albrecht Wacker · Jochen Kramer

Zusammenfassung: Vergleichsarbeiten (in einigen Bundesländern Lernstandserhebungen genannt) sollen innerhalb der aktuellen Umsteuerungen des Bildungswesens den Lehrkräften ein Referenzsystem für die Entwicklung von Unterricht zur Verfügung stellen. In Baden-Württemberg wurden sie erstmals zum Ende des Schuljahrs 2005/2006 verpflichtend implementiert. Die Lehrkräfte führten die Tests durch und korrigierten sie selbst. Anschließend erhielten sie eine Rückmeldung auf Schul- und Klassenebene, die in einer vorausgehenden landesweiten Pilotierung gewonnen wurde. In einer wiederholten Befragung wurde untersucht, welche intendierten Wirkungen (Orientierungsleistung) und nicht-intendierten Wirkungen (narrowing-the-curriculum-Effekt, curriculare Verengungen, Übungs- und Wiederholungserfordernisse) Lehrkräfte mit dem neuen Steuerungsinstrument der Vergleichsarbeiten verbinden. Befragt wurden Realschullehrkräfte in Baden-Württemberg. Es zeigte sich in der Befragung *vor* Einführung der Vergleichsarbeiten (2004; $n_{t1}=914$), dass die Lehrkräfte sowohl intendierte Wirkungen erwarteten als auch nicht-intendierte Wirkungen. Vier Jahre *nach* der Einführung der Vergleichsarbeiten (2009; $n_{t2}=734$) wurden die eingetretenen Wirkungen eingeschätzt. Sie fielen in allen Wirkungsdimensionen signifikant geringer aus als vor der Einführung erwartet wurde. Erfreulich ist, dass Lehrkräfte den befürchteten narrowing-the-curriculum-Effekt nicht als bedeutsam einschätzten. Dagegen sahen die Lehrkräfte in den Vergleichsarbeiten auch keine nennenswerten Orientierungshilfen zur Unterrichtsplanung und Leistungsbeurteilung. Vergleichsarbeiten wurden damit von den befragten Lehrkräften nicht als ein Referenzsystem für ihre Unterrichtsentwicklung angesehen.

Schlüsselwörter: Vergleichsarbeiten · Unterrichtsentwicklung · Realschule · Narrowing-the-curriculum · Längsschnittbefragung · Sekundarstufe

Online publiziert: 09.11.2012

Prof. des. Dr. A.Wacker (✉)
Institut für Erziehungswissenschaft, Pädagogische Hochschule Heidelberg,
Keplerstraße 87, 69120 Heidelberg, Deutschland
E-Mail: wacker@ph-heidelberg.de

Dr. J. Kramer (✉)
Institut für Erziehungswissenschaft, Empirische Bildungsforschung und
Pädagogische Psychologie (EBPP), Universität Tübingen,
Europastr. 6, 72072 Tübingen, Deutschland
E-Mail: jochen.kramer@uni-tuebingen.de

Standard Comparative Testing in Baden-Württemberg – On the judgement of teachers before and after implementation

Abstract: Standard comparative tests are meant to provide a reference system within the framework of educational reforms for development of school lessons. Compulsory comparative testing was introduced in Baden-Württemberg at the end of the school year 2005/2006. The teachers carried out the tests and corrected them themselves. Afterwards they received feedback at school and class levels, which was based on a prior statewide pilot study. The repeated survey had the aim of investigating which intended effects (reference attainment for orientation) and non-intended effects (narrowing-the-curriculum, pre-testing and exercises) teachers associated with this new instrument. Teachers of technical secondary schools (*Realschulen*) in Baden-Württemberg were surveyed. In the survey before the introduction of standard comparative testing (2004; $n_{tl}=914$), teachers expected both intended and non-intended effects. Four years after their introduction (2009; $n_{tl}=734$), respondents were asked to estimate the effects of standard comparative tests. The effects—in every dimension—were judged to be significantly less than had been expected before their introduction. It is pleasing that the teachers did not judge the anticipated narrowing-the-curriculum effects to be significant. However, they also did not see the instrument as a noteworthy orientation help for planning and assessing lessons. Standard comparative tests were not seen by the respondents to provide a reference for new lesson developments.

Keywords: Lesson development · Longitudinal surveys · Narrowing-the-curriculum · Realschule · Standard comparative tests · Technical secondary school

1 Einleitung

Beginnend in den 1990er-Jahren und verstärkt durch die internationalen Schulleistungsstudien hat in Deutschland eine Umsteuerung des Bildungswesens begonnen. Im Fokus dieser sogenannten *Neuen Steuerung* stehen die kurzfristigen und langfristigen Erträge (*output* und *outcome*) des Bildungssystems (Maag Merki 2010). Zur Sicherstellung der Erträge wurden rechenschaftsbasierte Steuerungssysteme eingeführt: auf der Makroebene des Bildungssystems Bildungsmonitoring, auf der Mesoebene Schulinspektionen und auf der Mikroebene Vergleichsarbeiten – auch Kompetenztests oder Lernstandserhebungen genannt. Wenngleich Vergleichsarbeiten auf der Unterrichtsebene angesiedelt sind, werden ihre Ergebnisse auch zur Steuerung auf der Schul- und Systemebene genutzt (Bonsen und von der Gathen 2004). Sie sind deshalb ein Kernelement der Neuen Steuerung, mit dem große Wirkhoffnungen verbunden sind. Konkret handelt es sich um Klassenarbeiten, die bundeslandweit zentral gestellt und zum gleichen Zeitpunkt geschrieben werden. Ihr Ziel ist es, den Lehrkräften einen Referenz- und Orientierungsrahmen für ihr unterrichtliches Handeln zu bieten, der über einen Vergleich der erzielten Testwerte auf individueller und Klassenebene mit den landesweiten Ergebnissen gewonnen werden soll. Zwischen den Bundesländern unterscheiden sich die Steuerungskonfigurationen und Durchführungspraxen der Vergleichsarbeiten teilweise erheblich. Welche Wirkungen Vergleichsarbeiten entfalten können, wird maßgeblich von der Art ihrer Implementation beeinflusst. Beispielsweise spielt die Art, ob sie mit Sanktionen verbunden sind oder nicht, eine bedeutsame Rolle.

Im vorliegenden Artikel wird die Implementation der Vergleichsarbeiten an Realschulen in Baden-Württemberg fokussiert. Zunächst wird hierzu die Konzeption und Durchführung der Vergleichsarbeiten in Baden-Württemberg beschrieben. Danach referieren wir Forschungsbefunde aus Ländern, die mit und ohne Sanktionssystemen bei Testungen arbeiten, um daran mögliche Wirkungen und Nebenwirkungen präziser herauszuarbeiten. Im Empirieteil werden die Wirkhoffnungen, die Lehrkräfte mit der Einführung der Vergleichsarbeiten verbinden, mit einer Trendstudie untersucht und im Diskussionsteil die Befunde abschließend zusammenfassend bewertet.

2 Zur Umsetzung der Vergleichsarbeiten in der Sekundarstufe in Baden-Württemberg

Zur Implementation und Durchführung der Vergleichsarbeiten in Baden-Württemberg liegt bislang keine detaillierte und umfassende Beschreibung vor. Das Land versuchte als erstes Bundesland eine systematische Output-Evaluation auf den Weg zu bringen und entwickelte dazu eigene Bildungsstandards unabhängig und vor den Nationalen Standards (Gördel 2008; Maier 2009). Implementiert wurden die Bildungsstandards und Vergleichsarbeiten in Baden-Württemberg gemeinsam mit weiteren Evaluationselementen der Outputsteuerung in einer umfassenden Bildungsreform zum Schuljahr 2004/2005, in die alle allgemeinbildenden Schulen einbezogen waren. Anders als in den bisherigen Lehrplänen des Bundeslandes wurden die Bildungsstandards dabei für mehrjährige Zeiträume ausgewiesen, z. B. für die Klassenstufen 5/6 oder 7/8, was den Lehrkräften die (neue) Aufgabe auferlegte, die zu vermittelnden Kompetenzen und Inhalte innerhalb der zweijährigen Zeiträume selbstständig zu verteilen und eine jahrgangsübergreifende Perspektive bei der Planung einzunehmen (Drieschner 2009). Mit der Bildungsreform wurde ebenfalls eine sogenannte Kontingentstundentafel eingeführt, in der für jede Schulart die Jahreswochenstunden festgelegt sind und die den Schulen eine Konzentration von Stunden in einzelnen Klassenstufen, beispielsweise in den Testjahrgängen, ermöglichte.

Nachdem die Lehrkräfte zwei Jahre lang nach den baden-württembergischen Standards unterrichtet hatten, wurden zum Ende des Schuljahrs 2005/2006 die ersten Vergleichsarbeiten flächendeckend geschrieben. Sie basierten auf vorausgehenden Entwicklungen des Landesinstituts, die 1999 begannen, und freiwilligen Probedurchgängen der Schulen (Sikorski 2003). In den ersten drei Durchgängen bis zum Schuljahr 2007/2008 wurden die Vergleichsarbeiten in den Klassenstufen 6 und 8 jeweils am Ende des Schuljahres geschrieben. Eine Änderung erfolgte zum Schuljahr 2009/2010. Ab diesem Zeitpunkt wurden die Arbeiten zu Beginn des Schuljahrs für die Klassenstufen 7 bis 9 angesetzt. Um eine Testhäufung zum Ende der Klassenstufe 6 bzw. 8 und zum Beginn der Klassenstufe 7 bzw. 9 zu vermeiden, entfielen die Testungen im Schuljahr 2008/2009.

Während die Fächer Mathematik und Deutsch feste Kernbestandteile der Testungen in beiden Klassenstufen waren, konnte die Einzelschule in den ersten drei Durchgängen einen weiteren Bereich auswählen: entweder den Fächerverbund EWG (=Erdkunde – Wirtschaftskunde – Gemeinschaftskunde) oder das Fach Geschichte. Zum Schuljahr 2008/2009 entfiel die freie Wahl dieses dritten Prüfungsfaches. Stattdessen wurde die

Prüfung der Pflichtfremdsprache (überwiegend das Fach Englisch, entlang der Rheinschiene das Fach Französisch) in Klassenstufe 9 eingeführt.

Innerhalb der Fächer wechselten die abgeprüften Schwerpunktbereiche und Dimensionen zyklisch über die Testzeitpunkte hinweg.[1] Auch die Itemformate zeigten innerhalb der Fächer jeweils fachspezifische Ausprägungen. Im Fach Englisch waren dies beispielsweise das Finden von Wörtern im Text, die Bearbeitung von Lückentexten, das Fertigschreiben von begonnenen Sätzen, die Beantwortung von Fragen und das Verfassen freier Texte innerhalb eines eng gesteckten Rahmens. Im Fächerverbund EWG dagegen galt es unter anderem Lösungen zuzuordnen, Maßstabsberechnungen und Begriffsbestimmungen durchzuführen sowie Himmelsrichtungen in Pfeildiagrammen zu ergänzen, aber auch freie Texte und Begründungen zu verfassen.

Gegenüber den gebräuchlichen Klassenarbeiten, die von den Lehrkräften verwendet werden, sind Unterschiede vor allem in der Leistungsbeurteilung, aber auch in den Formen der verwandten Aufgaben zu erkennen.[2] So unterscheidet sich z. B. die Bepunktung (in Klassenarbeiten sind auch halbe Punkte möglich, in Vergleichsarbeiten nur ganze Punkte; Sikorski 2003) und das Aufgabenformat (in Klassenarbeiten auch offene Fragen, in Vergleichsarbeiten geschlossene Fragen; Rothe 2003). Insgesamt haben die Aufgabenformate der Vergleichsarbeiten stärker den Charakter von Testaufgaben,[3] die einen Transfer erlernter Kompetenzen beinhalten, weniger den von Lernaufgaben. Rothe (2003) beziffert die Korrelationen zwischen den durch Klassenarbeiten gewonnenen Noten für die Hauptfächer und die entsprechenden Vergleichsarbeiten auf $r = 0{,}48$ bzw. $r = 0{,}64$ (ohne Angabe des konkreten Faches für die Korrelation) und interpretiert dies als eine sehr eingeschränkte Validität der Arbeiten. Weiterhin berichtet Rothe, dass in dieser Anfangszeit die Konventionswerte der Trennschärfen und Reliabilitätswerte für die Vergleichsarbeiten (die damals noch „Jahrgangsarbeiten“ hießen) nicht immer erreicht wurden. Er führte die unbefriedigenden Validitäts- und Reliabilitätskennwerte darauf zurück, dass die Zeugnisnoten der Schülerinnen und Schüler insgesamt nur eine geringe Varianz aufwiesen. So erhielt z. B. im Fach Deutsch die Hälfte die Note „befriedigend“.

Von zentraler Bedeutung für die Wirksamkeit der Vergleichsarbeiten ist die Rückmeldung, welche die Lehrkräfte und Schulen erhalten und der hierbei zugrundeliegende Vergleichsmaßstab (vgl. Fiege et al. 2011). Die Lehrkräfte in Baden-Württemberg erhielten hierzu Vergleichswerte, die aus einer vorab erfolgten Pilotierung der Aufgaben gewonnen wurden. Die Pilotierung fand jeweils etwa ein Jahr vor dem Testeinsatz der Aufgaben in ausgewählten Schulen des ganzen Landes statt. Aus ihren Ergebnissen heraus werden drei Leistungsgruppen für die Rückmeldungen an die Lehrkräfte festgelegt: eine Leistungsgruppe, welche die unteren 25 % umfasst, eine Leistungsgruppe mit den mittleren 50 % und eine Leistungsgruppe mit den oberen 25 %.[4] Den Rückmeldungen liegen somit ausschließlich gruppenbezogene Normen unter Verwendung der sozialen Bezugsnorm zugrunde. Bei der Zusammenstellung der Vergleichsarbeiten werden dabei die Aufgaben anhand der Pilotierungsergebnisse so ausgewählt, dass ein „mittlerer“ Schüler etwa 50 % der Aufgaben löst, was der Note 3 = „befriedigend“ entspricht.

Abweichend von anderen Bundesländern (z. B. Hamburg), aber affin zu allen Ländern, die sich an VERA beteiligen, wurden die Vergleichsarbeiten durch die Lehrkräfte selbst korrigiert und mit einer Note, welche in die Leistungsbeurteilung einfloss, versehen.[5] Eine Rückmeldung der Ergebnisse an die Lehrkräfte erfolgte leistungsgruppenbezogen

auf der Aufgabenebene, der Ebene der Schülerinnen und Schüler sowie auf der Klassenebene. Die Lehrkräfte erhielten Informationen sowohl in tabellarischer als auch in grafischer Form (vgl. für Beispiele Anhang A und B). Auf der Aufgabenebene wurden die Lösungshäufigkeiten in Prozent ausgewiesen. Differenzen von 20 % und mehr wurden als bedeutsam gekennzeichnet. In den jährlich ausgewiesenen Schwerpunktbereichen wurde darüber hinaus grafisch dargestellt, wie sich die Klasse über die Leistungsgruppen verteilt. Auf Schülerebene wurde die Anzahl der richtig gelösten Items und der Test-Score (Summe der gewichteten richtig gelösten Items) ausgewiesen; außerdem wurde angegeben, welcher Leistungsgruppe die einzelnen Schülerinnen und Schüler angehören und welche Noten ihnen zuzuordnen sind. Die Auswertung auf Klassenebene schließlich stellte die Mittelwerte einer Schulklasse den landesweiten Vergleichswerten gegenüber (in den Einzelitems, Schwerpunktbereiche und erzielten Leistungsgruppen). Vergleichswerte für weitere Subpopulationen, beispielsweise variierende Referenzgruppen, welche die soziale Belastung von Schulen quantifizieren, wurden darüber hinaus nicht gegeben. In Baden-Württemberg bezogen sich die Vergleichswerte ausschließlich auf bestimmte Schwerpunktbereiche eines bestimmten Fachs in einer bestimmten Klassenstufe und Schulart.

Die Schulleitung erhielt als Rückmeldung eine Kompilation aller Klassenrückmeldungen. Den Schulen war es freigestellt, den Eltern eine Rückmeldung der Tests zukommen zu lassen und dabei die Daten der Schülerebene zu berücksichtigen.

Mit den Vergleichsarbeiten in Baden-Württemberg sind keine Konsequenzen für Schule und Lehrpersonen verbunden, wie dies beispielsweise in den angelsächsischen Ländern der Fall ist (s. Abschn. 4); der Umgang mit Ergebnissen der Tests ist ausschließlich in die Verantwortung der Einzelschulen und Lehrkräfte gelegt und variiert zwischen den Schulen und Lehrkräften. Dem umfangreichen Begleitmaterial, das die Lehrkräfte zur Durchführung erhalten, wurde unterstützend eine Handreichung mit dem Titel „Umgang mit den Ergebnissen im Rahmen der Selbstevaluation der Schulen“ beigegeben (aktuelle Fassung: Landesinstitut für Schulentwicklung 2010). Darin wird eine Nutzung der Daten entlang des Dreischritts „Analyse der Ergebnisse“, „Interpretation der Ergebnisse“ und „Konsequenzen ziehen“ vorgeschlagen. Zur Analyse wird darin beispielsweise ausgeführt, wie die Lehrkräfte auftretende Differenzen (zwischen den von ihnen erzielten Ergebnissen und dem Landesdurchschnitt) in ihrer Richtung erkennen und diese mit den eigenen Erwartungen vergleichen können. Hinsichtlich der Interpretation der Daten sind Denkhilfen für Erklärungsansätze auf verschiedenen Ebenen angedacht (einzelne Schüler, Klasse, Unterricht/Fach, Schule), welche der Intention verpflichtet sind, dass die Lehrkräfte Ziele für ihr weiteres Arbeiten generieren – mögliche Beispiele sind vorgedacht. Über diese Vorschläge hinausgehende externe Unterstützungen zur Interpretation der Ergebnisse, beispielsweise Fortbildungen für Lehrkräfte, wurden nicht angeboten. Die dargelegte Konzeption der Arbeiten ermöglicht es den Schulen, variierende Strategien im Umgang mit den Ergebnissen zu praktizieren.

3 Zu den intendierten Funktionen der Vergleichsarbeiten in Baden-Württemberg

Welche Funktionen und welcher angedachte „Mehrwert", so bleibt zu fragen, sind mit diesem für Lehrkräfte recht aufwendig durchzuführenden Verfahren im Rahmen der neuen Steuerungsphilosophie intendiert? Mit Vergleichsarbeiten ist im Allgemeinen das Ziel verbunden, schulische Arbeit überprüfbar zu machen. Ihnen kommen deshalb differente Funktionen auf den verschiedenen Ebenen des Bildungssystems zu (vgl. Tab. 1).

Auf der Mikroebene ist intendiert, den Lernstand von Klassen und von einzelnen Schülerinnen und Schülern in Bezug auf die Bildungsstandards zu überprüfen und damit förderdiagnostisches Wissen für die Lehrkräfte zu erzielen. Auf der Mesoebene der Schule werden die Ergebnisse der Vergleichsarbeiten als eine datenbasierte Grundlage zur Schulentwicklung herangezogen und auf der Makroebene des Systems sind die gewonnenen Daten Teil des Bildungsmonitorings.

Im Kern fokussieren Vergleichsarbeiten im neuen Steuerungskonzept die Mikroebene des Unterrichts (Oelkers und Reusser 2008). Auch in Baden-Württemberg steht die Überprüfung der Standarderreichung hierbei im Mittelpunkt: „Die Vergleichsarbeiten vermitteln den Lehrkräften, den Schülerinnen und Schülern sowie deren Eltern in bestimmten Fächern objektive Informationen über den individuellen Lernstand im Hinblick auf ausgewählte Bereiche der Bildungsstandards" (Landesbildungsserver Baden-Württemberg 2009). Nach dieser offiziellen Verlautbarung geben Vergleichsarbeiten den Lehrkräften Hinweise zur Unterrichtsplanung. Scholl (2009) konkretisiert die Aspekte dieser *Orientierungsfunktion* mit Bezug auf die Lehrkräfte in theoretischer Hinsicht und subsumiert die Unterrichtsplanung, die Leistungsbeurteilung im Allgemeinen und das Erkennen von Lernrückständen im Besonderen darunter. Dabei stellt sich die Frage, ob aus der Sicht von Lehrkräften die Vergleichsarbeiten in Baden-Württemberg diese intendierte Orientierungsfunktion zu erfüllen vermögen und welche weiteren Wirkungen sie damit ver-

Tab. 1: Generelle Funktionen der Vergleichsarbeiten

Ebene im Schulsystem	Funktionen der Vergleichsarbeiten
Makroebene System	Bildungsmonitoring und Bildungsberichterstattung als Information der Politik und Öffentlichkeit
	Objektiver und landesweiter Vergleichsmaßstab bei der Leistungsbeurteilung durch Lehrkräfte
Mesoebene Schule	Instrument zur Qualitätsentwicklung an Schulen im Rahmen der Fremd- und Selbstevaluation, damit auch Instrument der Schulentwicklung
	Kontrolle der Lehrkraft über Lehrplanumsetzung
	Informationen für die Schulaufsicht und für die Eltern
Mikroebene Unterricht	
Schüler	Information zum Lernstand für Schülerinnen und Schüler
	Zuweilen Funktion als Klassenarbeit
Lehrer	Diagnoseinstrument für Lehrkräfte
	Instrument zur Unterrichtsentwicklung

binden. Ein wesentlicher Forschungsbefund zu Instrumenten neuer Steuerung ist, dass mit ihrer Implementation auch vielfach nicht-intendierte oder sogar kontraintentionale Nebenwirkungen verbunden sind (Bellmann und Weiß 2009). Beispielsweise können Vergleichsarbeiten zu einer Konzentration auf die Kompetenzbereiche, Aufgabenformate und Fächer führen, die in den Vergleichsarbeiten selbst relevant sind. Dies kann zwar eine durchaus erwünschte Steuerungsfunktion sein, es kann jedoch auch dazu führen, dass wichtige, aber für die Vergleichsarbeiten weniger relevante Inhalte und Aufgabenformate vernachlässigt werden (*narrowing the curriculum*- und *teaching to the test*-Phänomone; vgl. z. B. Herrmann 2005, 2010; Künzli 2006; Plöger 2005; Regenbrecht 2005; Schirp 2006).

Bevor wir uns der Frage zuwenden, wie verbreitet die formulierten Erwartungen und Befürchtungen unter den Realschullehrkräften in Baden-Württemberg sind, berichten wir Befunde aus internationalen und nationalen Studien, die die Implementation und Wirkung von Vergleichsarbeiten thematisieren. Im Vergleich dazu werden die Charakteristika der Baden-Württembergischen Vergleichsarbeiten verdeutlicht und Forschungsbefunde zu Vergleichsarbeiten in Baden-Württemberg berichtet.

4 Forschungsbefunde zu Wirkungen von Vergleichsarbeiten aus anderen Staaten und Bundesländern

In der Forschung sind häufiger Befunde aus Staaten aufzufinden, die mit Sanktionssystemen (*high stakes*) arbeiten. Ergebnisse zu high-stakes-tests liegen insbesondere aus den USA vor. Diese wurden in den USA infolge der Publikation „A Nation at Risk“ (National Commission on Excellence in Education 1983) in den 1980er-Jahren implementiert. Eine nationale Ausweitung erfolgte 1994 mit dem *Improving America's School Act*, der die Bundesstaaten verpflichtete, sowohl *content* und *performance standards* als auch ein System der Leistungserhebung und Leistungsbewertung (*assessment*) für alle Schülerinnen und Schüler einzuführen. Eine nochmalige Bedeutungszunahme fand mit dem N*o Child Left Behind Act* (NCLB) von 2002 statt: Im NCLB-Konzept ist eine jährliche Steigerung der in Tests erzielten Punktwerte der Schulen anvisiert (*adequate yearly academic progress*) und für den Fall, dass der anvisierte jährliche Fortschritt nicht erreicht wird, sind abgestufte Sanktionsmaßnahmen für die Schulen vorgesehen, beispielsweise das Recht für Schülerinnen und Schüler, auf eine andere öffentliche Schule zu wechseln (Hursh 2005). Im Gegensatz zu den neuen Steuerungsinstrumenten in Deutschland wird damit in den USA ein starker Druck auf die Schulen ausgeübt.

Koretz et al. (1996a, 1996b) berichteten, dass die Lehrkräfte ihre Praxis an den Curricula und Tests ausrichteten und wie gewünscht beispielsweise vermehrt kooperative Lernformen und lebensnahe Bezüge der Inhalte betonten. Nach der Implementierung von NCLB berichteten Hamilton et al. (2007) aus einer Studie mit insgesamt 361 Lehrkräften, welche in den drei amerikanischen Bundesstaaten Kalifornien, Pennsylvania und Georgia durchgeführt wurde, dass die Lehrerinnen und Lehrer mit den implementierten Content Standards im Fach Mathematik vertraut waren und diese hilfreich für die Unterrichtsplanung fanden.

Stecher (2002) fasste den Forschungsstand der positiven Effekte aus den Testungen insgesamt dahingehend zusammen, dass die Tests zu erhöhten Anstrengungen der Lehrkräfte in der schulischen Praxis führten. Solch positive Befunde wurden aber auch kritisiert, weil sie mehr die Reaktion auf spezifische Testelemente abbilden und lernpsychologische und fachdidaktische Entwicklungen unzureichend berücksichtigen würden (z. B. Stecher und Barron 2001).

Welche darüber hinausgehenden Wirkungen der Tests wurden von der Forschung benannt? Empirisch gut abgesichert sind Befunde, die zeigten, dass Lehrkräfte verstärkt jene Fähigkeiten und jenes Wissen unterrichten, die mit den Testungen abgeprüft werden (z. B. Koretz et al. 1996a). Andere Studien berichteten von umfangreichen Vorbereitungen auf die Testungen selbst. Dazu gehörten die Betonung höherer kognitiver Leistungen, die Erhöhung der Motivation der Studierenden in Bezug auf die Tests und die Verwendung von Tests für Übungen (Koretz et al. 1996b). Auch qualitative Forschungen stützten diese Ergebnisse und zeigten die Bedeutung von Testvorbereitungen auf (z. B. Lipman 2004). In der bereits erwähnten Studie berichteten Hamilton et al. (2007) von Zeitausweitungen jener Fächer, die in den Tests abgeprüft wurden. Zuweilen wurden diese Zeitveränderungen für alle Schüler eingeführt, manchmal auch nur für die schwachen Schüler. In diesem Zusammenhang wies Hursh (2005) auf Fälle hin, in denen bestimmte Fächer überhaupt nicht mehr unterrichtet wurden. Dass zur Testvorbereitung häufig Unterrichtszeit umgewidmet wurde, ist ein gut abgesichertes Forschungsergebnis.

Unter high-stakes-Bedingungen zeigen sich so deutliche Effekte von Vergleichsarbeiten: Sie bieten eine Orientierungsfunktion für die Lehrkräfte und können zu narrowing-the-curriculum- und teaching-to-the-test-Effekten führen (Stecher 2002; Herman 2004). Forschungsmethodisch ist einschränkend zu erwähnen, dass zahlreiche Befunde auf Querschnittsbefragungen von Lehrkräften bzw. Schulleitern beruhen. Welche Änderungen der Unterrichtspraxis und Inhalte infolge der Tests tatsächlich bestehen und ob diese überhaupt auf das implementierte Accountability-System zurückgeführt werden können, ist unklar (z. B. McDonnell und Choicer 1997).

Die Frage, ob sich Wirkungen von Vergleichsarbeiten auch dann zeigen, wenn sie nicht mit Sanktionen für die Schulen und Lehrkräfte verbunden sind (*low-stakes*), lässt sich anhand von Studien prüfen, die in den Niederlanden und Schweden durchgeführt wurden. In den Niederlanden wurde seit etwa 2000 ein umfassendes Accountability-System implementiert, welches auf Sanktionsmechanismen verzichtet. Verpflichtende Tests, zu denen auch freiwillige Tests treten können, werden zum Ende der Grundschulzeit und zu den Abschlussprüfungen eingefordert. Eine Besonderheit der Implementation in den Niederlanden ist, dass die Schulen einen Entscheidungsspielraum besitzen, welchen Test sie anwenden möchten; über 80 % der Schulen nutzen hierbei ein von Cito bereitgestelltes Instrument (Berkemeyer und van Holt 2012). Die Testergebnisse werden in Form von schülerspezifischen Berichten rückgemeldet (Oelkers und Reusser 2008). Eine umfassende Studie zu Accountability-Effekten in den Niederlanden liegt in einer Studie von Chorny und Webbink (2010) vor, in die Daten von 60.000 Schülerinnen und Schülern aus 600 Schulen eingeflossen sind. Chorny und Webbink untersuchten das strategische Verhalten der Schulen zur Vorbereitung der Testungen. Sie nahmen an, dass das Verwenden des Vorjahrestests zur Übung eine einseitige Steigerung der fachlichen Kompetenzen evoziert, nicht aber der überfachlichen Kompetenzen. Sie fanden das Gegenteil: 60 % des

Kompetenzzuwachses waren auf die überfachlichen Kompetenzen zurückzuführen, ein teaching-to-the-test-Effekt war nicht nachzuweisen.

Schweden sieht seit 1962 einen limitierten Einsatz von Prüfungen vor: Diagnostische Prüfungen in der 2. Klasse und Fachprüfungen in der 5. Klasse sind freiwillig, lediglich die Prüfungen in der 9. Klasse verpflichtend. Ein Einfluss der Tests auf die Lehrkräfte und deren Praxis wird in der Literatur zwar angenommen (z. B. Erickson und Lander 2007), abgesicherte Evidenzen dafür sind aber nicht berichtet worden. Dies kann seinen Grund darin haben, dass die Testergebnisse nicht zwingender Bestandteil der erteilten Note sind und die Lehrkräfte einen Ermessensspielraum bei der Leistungsbeurteilung besitzen (Waldow 2005). Insgesamt scheint dem Thema in Schweden eine nachgeordnete Relevanz zuzukommen (vgl. z. B. die Ausführungen von Carlgren 2009), dies vermutlich auch deshalb, weil die ausdrückliche Hauptfunktion der schwedischen Tests nicht in einer Referenzfunktion für die Lehrkräfte liegt, sondern darin, ein einheitliches Beurteilungsniveau über das ganze Land sicherzustellen (Waldow 2005).

Auch in den Ländern der Bundesrepublik Deutschland wurde die Implementation von flächendeckenden Tests im Rahmen Neuer Steuerung nicht mit Sanktionssystemen gekoppelt. Ein Schwerpunkt der Forschung zu Vergleichsarbeiten in Deutschland bildete vor allem die Befragung von Lehrkräften zu Rezeption und Nutzen von Tests, in denen auch die Frage nach den orientierenden Wirkungen der Tests berücksichtigt wurde. Helmke (2004) fand in einer Befragung zu den erfolgten Rückmeldungen aus der MARKUS-Studie, dass die Mehrheit der Lehrkräfte keine unterrichtsbezogenen Maßnahmen ergriff. Groß Ophoff et al. (2006) berichteten aufgrund einer standardisierten Befragung zum Projekt VERA, dass die Lehrkräfte Vergleichsarbeiten als aufwendig empfanden und Veränderungen des Unterrichts die Ausnahme darstellten. Die Nützlichkeit und die Beschaffenheit der Rückmeldungen sowie die Überzeugungen der Lehrkräfte untersuchten z. B. Kuper und Hartung (2007) sowie Schneewind und Kuper (2009). In diesen Studien konnten bislang kaum Hinweise darauf gefunden werden, dass Lehrkräfte aufgrund der Rückmeldungen ihre didaktisch-methodische Handlungsweisen im Unterricht änderten. Eine Begleitbefragung zu den Thüringer Kompetenztests zeigte, dass die Lehrkräfte den Nutzen der Tests vor allem in der Leistungsdiagnostik und weniger in der Weiterentwicklung von Unterricht sahen (Nachtigall und Jantowski 2007).

5 Forschungsbefunde zur Implementation der Vergleichsarbeiten in Baden-Württemberg

Erste empirische Ergebnisse zur Wirkung von Vergleichsarbeiten in Baden-Württemberg liegen von Maier und Rauin (2006) vor. Sie befragten Lehrkräfte in allen Schulformen der Sekundarstufe I in Baden-Württemberg zur Akzeptanz zentraler Tests mit einem standardisierten Instrument ein Jahr nach Einführung der Reform. Zu diesem Zeitpunkt, hatten in der Sekundarstufe I noch wenige Lehrkräfte Erfahrung mit den Vergleichsarbeiten. Ausgewertet wurden 1.294 Fragebögen (davon 605 aus Hauptschulen, 405 aus Realschulen und 284 aus Gymnasien). Lehrkräfte, die die Vergleichsarbeiten bereits 2003 und 2004 freiwillig erprobten, akzeptierten sie eher als diejenigen Lehrkräfte, die keine Vorerfahrung mit ihnen besaßen. Direkt im Anschluss an die flächendeckende Imple-

mentation in Baden-Württemberg erhob Maier (2008a; 2009) die allgemeine Akzeptanz zu den Arbeiten erneut und verglich schulformbezogen ihre curriculare Validität, ihre förderdiagnostische Nutzung sowie ihre Orientierungsfunktion für die Unterrichtsgestaltung und ihre selektionsdiagnostische Nutzung mit der ersten Befragung. In diese zweite Befragung gingen die Urteile von 307 Lehrkräften ein (Hauptschulen 113, Realschulen 81, Gymnasien 113).

Insgesamt stellte er eine deutliche Abnahme der Akzeptanz von Vergleichsarbeiten fest und führt aus, dass Gymnasial- und Realschullehrkräften den Vergleichsarbeiten zunächst kritischer gegenüberstanden als Hauptschullehrkräfte. Der Anteil der Lehrkräfte mit skeptischer Einstellung nahm in der zweiten Befragung zu. Vor allem die Hauptschullehrerinnen und -lehrer korrigierten ihre zunächst positive Einschätzung, dass mit Vergleichsarbeiten die Individualförderung forciert werden könnte, zum zweiten Zeitpunkt (Maier 2009).[6] Zusammenfassend weisen die Ergebnisse für Baden-Württemberg einerseits auf eine Orientierungsleistung, die von den flächendeckenden Assessments für die Lehrkräfte ausgeht, hin, andererseits darauf, dass Wirksamkeit und Akzeptanz von Vergleichsarbeiten im Zeitverlauf abnahmen. Die Befürchtung von curricularen Verengungen wird von den Lehrkräften nicht geteilt und insgesamt keine „extreme Deformation des Unterrichts“ (Maier 2010) festgestellt. Auch Hinweise darauf, dass schulische Bildungsziele eingeschränkt werden, konnten nicht aufgefunden werden (Maier 2009). Zur Interpretation der Daten aus den gezogenen Querschnitten gibt der Autor einschränkend zu bedenken, dass die Befunde auf schulformspezifisch kleinen Stichproben mit möglichen Selektionseffekten beruhen und sie nur quasi-längsschnittlich zu betrachten sind (Maier 2009). Diesen Einschränkungen versucht die vorliegende Studie mit einer schulformspezifisch großen Stichprobe unter Verwendung eines echten Längsschnittdesigns entgegenzutreten.

6 Zu Forschungsdesideraten und zur Fragestellung

Die Befunde aus den Ländern mit high-stakes-tests weisen mit hoher Evidenz auf, dass dort die Tests eine orientierende Funktion besitzen, aber auch mit curricularen Verengungen, Testvorbereitungen und Zeitumwidmungen einhergehen. Dagegen zeigen sich diese Ergebnisse aus Ländern, die nicht mit Sanktionen arbeiten (low-stakes-tests), nicht im gleichen Maß und in gleicher Ausprägung. Diese Variation der Befunde ist vermutlich auf die differenten Funktionen, die mit den Tests in unterschiedlichen Steuerungskonfigurationen verbunden sind, zurückzuführen. Welche Erwartungen und Befürchtungen mit der Einführung der Vergleichsarbeiten bei Lehrkräften an Realschulen in Baden-Württemberg verbunden sind, wird mit der vorliegenden Studie untersucht. Folgende Fragen werden betrachtet:

Die erste *Frage zur Orientierungsfunktion* wendet sich dem Bereich der erwünschten Effekte zu: Als wie hilfreich bewerten Lehrkräfte die Vergleichsarbeiten zur Unterrichtsplanung, zur Leistungsbeurteilung und zum Erkennen von Lernrückständen vor und vier Jahre nach ihrer Einführung? Infolge der Tatsache, dass der Umgang mit den Vergleichsarbeiten weitestgehend den Einzelschulen überantwortet wurde, ist zu erwarten, dass eine große Streuung im Lehrerurteil aufgefunden wird. Mit Verweis auf die bisherigen For-

schungsbefunde aus Baden-Württemberg ist zu vermuten, dass zum zweiten Messzeitpunkt die unterrichtliche Orientierungswirkung weniger stark eingeschätzt wird, als zum ersten Messzeitpunkt erwartet.

Die beiden weiteren Fragen gelten möglichen unerwünschten Effekten der Testungen. Mögliche curriculare Verengungen fassen wir im Begriff „narrowing-the-curriculum". Unter ihm verstehen wir die inhaltliche und methodische Ausrichtung des Unterrichts auf die Kompetenzbereiche, die in den Vergleichsarbeiten abgeprüft werden. Die zweite Frage zu dieser möglichen *narrowing-the-curriculum-Wirkung* lautet: Führen Vergleichsarbeiten aus Sicht der Lehrkräfte vor und vier Jahre nach der Einführung zu narrowing-the-curriculum-Phänomenen, die sich in einer Konzentration auf die Kompetenzbereiche, Aufgabenformate und Fächer zeigen, die in den Vergleichsarbeiten relevant sind? Wie zu ersehen war, liegt in der Konzeption der Arbeiten und Rückmeldepraxis die Möglichkeit variierender Strategien im Umgang mit den Ergebnissen auf der Schulebene (bis hin zu schuleigenen Sanktionierungsmechanismen) begründet. Die in Baden-Württemberg implementierte Kontingentstundentafel bietet auf der Schulebene die organisatorische Möglichkeit, Stunden in den abgeprüften Fächern der Testjahrgänge zu konzentrieren und andere Fächer weniger zu bedienen („Fachvernachlässigung"). Aufgrund der Anlage der Vergleichsarbeiten als low-stakes-tests erwarten wir entsprechend des Forschungsstandes geringe bis keine curricularen Verengungen zu beiden Messzeitpunkten, weil auf der Systemebene keine Sanktionsmechanismen folgen. Über einzelschulische Strategien und Sanktionssysteme liegen bislang, besonders im Hinblick auf Änderungen im Zeitverlauf, keine Befunde vor.

Die dritte Frage bezieht sich auf die *Übungs- und Wiederholungserfordernisse:* Führen die Vergleichsarbeiten aus Sicht der Lehrkräfte vor und vier Jahre nach der Einführung dazu, dass die Lehrkräfte im Unterricht mehr Üben und Wiederholen? Einerseits führt die Tatsache, dass in den Vergleichsarbeiten die Inhalte aus zwei Klassenstufen – als Klassenarbeitet benotet – abgefragt werden, zu dieser Vermutung, weil anzunehmen ist, dass Lehrkräfte in den abgeprüften Fächern die unterrichteten Themen mit ihren Schülern wenigstens in Grundzügen rekapitulieren. Zudem wechseln – wie oben bereits dargestellt – die jeweils getesteten Schwerpunktbereiche in den feststehenden Testfächern. Somit ist zu vermuten, dass die Lehrkräfte die in den letztjährigen Tests nicht enthaltenen Schwerpunktbereiche nochmals rekapitulierend aufgreifen. Qualitative Befunde aus den Studien Maiers unterstützen diese Annahme (Maier 2009), deren quantitative Überprüfung steht noch aus. Andererseits ist jedoch zu bedenken, dass die Vergleichsarbeiten als Testaufgaben konzipiert sind, die einen Kompetenztransfer beinhalten. Bislang ist unklar, ob und inwiefern die Lehrkräfte ihre Schülerinnen und Schüler mit vermehrtem Üben und Wiederholen hierauf vorbereiten können. Ist dies aus Sicht der Lehrkräfte nicht der Fall, so ist davon auszugehen, dass diese nicht vermehrt Üben und demzufolge auch keine vermehrten Übungs- und Wiederholungserfordernisse berichten. Wir nehmen beide Argumentationslinien berücksichtigend an, dass die aufgeworfene Frage zum ersten Messzeitpunkten überwiegend bejaht wird, weil den Lehrkräften die Funktion der Arbeiten und die Charakteristik ihrer Aufgaben noch zu wenig bekannt war; zum zweiten Messzeitpunkt rechnen wir mit Verweis auf das Argument zum Kompetenztransfer mit einer geringeren Einschätzung durch die Lehrkräfte.

7 Methode

Die hier berichtete Studie stellt einen auf die Vergleichsarbeiten bezogenen Teilaspekt eines Forschungsprojektes dar, dessen größerer Themenzusammenhang die Rezeption der Bildungsstandards in Baden-Württemberg aus Sicht der Realschullehrkräfte umfasste (Wacker 2008). Das Projekt wurde an der Pädagogischen Hochschule Ludwigsburg begonnen und an der Universität Tübingen fortgeführt. Auf der Grundlage qualitativer Lehrerbefragungen wurde ein standardisiertes Befragungsinstrument generiert. Die erste Befragung wurde im Herbst 2005 durchgeführt, d. h. ein Jahr nach Implementation der Bildungsreform in Baden-Württemberg und zugleich ein Jahr vor dem ersten flächendeckenden Einsatz der Vergleichsarbeiten in diesem Bundesland. Die zweite Befragung fand 2009 statt. Bis dahin konnten die Lehrkräfte in vier Durchgängen Erfahrungen mit der Durchführung der Vergleichsarbeiten sammeln. Eine Vorstudie konnte aufzeigen, dass mit der Bildungsreform an den Realschulen größere programmatische und organisatorische Veränderungen einhergingen als an den anderen Schulformen. Aus diesem Grund wurde die Umsetzung der Reform an Realschulen fokussiert. Die hier referierten Daten beziehen sich ausschließlich auf benotete Vergleichsarbeiten, die als summative Leistungsmessungen geschrieben wurden. (Mittlerweile wurde diese Praxis zugunsten formativer Leistungsmessungen, deren Ergebnisse nicht in die schulische Leistungsbeurteilung einfließen, geändert.)

7.1 Stichprobe und Rücklauf

Vom Statistischen Landesamt wurde im März 2005 eine Stichprobe von 101 der 427 staatlichen Realschulen in Baden-Württemberg gezogen und alle 3.373 Lehrkräfte zur Befragung eingeladen, die an diesen Schulen unterrichteten. Private Realschulen wurden nicht in die Befragung einbezogen, weil davon ausgegangen werden konnte, dass kein direkter curricularer Vergleich zwischen Schulen in nicht-staatlicher Trägerschaft und staatlichen Schulen möglich war. Die Zahl angeschriebener Probanden umfasste ein Viertel der zu diesem Zeitpunkt unterrichtenden 13.510 Lehrkräfte in Voll- und Teilzeit (Wacker 2008). Referendare und kirchliche Religionslehrkräfte wurden nicht befragt.

Bei der Ziehung wurden drei Variablen zur Stratifizierung verwendet, um eine möglichst repräsentative Abbildung der Gesamtheit der Realschulen Baden-Württembergs in der Stichprobe zu erreichen: *Schulgröße* (<400, 400–600, >600 Schülerinnen und Schüler), *Raumstruktur* nach dem Landesentwicklungsplan (Verdichtungsraum, Randzonen und Verdichtungsbereiche, ländlicher Raum; vgl. Innenministerium Baden-Württemberg 1984) und die *Verwaltungsuntereinheiten* (Regierungspräsidien Freiburg, Karlsruhe, Stuttgart, Tübingen). Bei Verbundschulen, die dadurch gekennzeichnet sind, dass eine Haupt- und Realschule unter einer gemeinsamen Schulleitung in einem Haus zusammen sind, wurde die Zuordnung der Lehrkräfte zur jeweiligen Schulform über einen Anruf bei den Schulleitungen erfragt. Die Teilnahmequote zum ersten Messzeitpunkt betrug 27,1 % (n_{tl} = 914), ein geringer, aber für diesen Gegenstandsbereich nicht unüblicher Wert (vgl. z. B. Beer 2006; Böttcher und Dicke 2008; Vollstädt et al. 1999).

Zum zweiten Messzeitpunkt im Herbst 2009 wurden alle 101 Realschulen, die 2005 gezogen wurden, um erneute Teilnahme gebeten. 86 Schulen erklärten sich zur erneu-

ten Teilnahme bereit; eingeladen wurden alle 2.902 Lehrkräfte dieser Schulen (20,4 % aller 14.234 Voll- und Teilzeitlehrkräfte an Realschulen in Baden-Württemberg zu T2[7]). Die Teilnahmequote betrug 25,3 % (n_{t2} = 734). 141 Lehrkräfte aus 60 Schulen nahmen an beiden Befragungen teil. Dies ermöglichte es, eine Substichprobe längsschnittlich zu betrachten.

Infolge der Rücklaufquoten von lediglich etwas über 25 % ist die Frage zu stellen, ob Verzerrungen vorliegen. Um dies abzuschätzen, wurden verschiedene Analysen durchgeführt:

a. Die Stichproben-Daten wurden mit den Statistiken des Landesamtes verglichen: Hinsichtlich Alter und Geschlecht entsprechen die Stichproben in etwa den Populationen (vgl. Tab. 2).
b. Die Antworten der Lehrkräfte von Schulen, bei denen ein hoher Rücklauf vorliegt, wurden mit den Antworten derjenigen verglichen, die an Schulen mit geringem Rücklauf unterrichteten. Würden sich hier Abweichungen zeigen, könnte dies ein Hinweis auf Verzerrungen durch die Rücklaufquote sein. Die Einschätzungen der Vergleichsarbeiten durch die Lehrkräfte von Schulen mit hohem und geringem Rücklauf unterschieden sich jedoch in allen eingesetzten Items nicht signifikant voneinander (Gruppenbildung durch Mediansplit; vgl. Wacker 2008).
c. Die Höhe des Rücklaufs zu beiden Messzeitpunkten war ungefähr gleich hoch (27,1 % und 25,3 %).
d. Eine Teilstichprobe nahm zu beiden Messzeitpunkten an der Befragung teil. Betrachtet man diese Längsschnittstichprobe, lassen sich die Ergebnisse, die auf den Querschnittstichproben basieren, replizieren (vgl. Ergebnisse, Abschn. 8). In den vorliegenden *quantitativen* Daten selbst konnten somit keine Hinweise auf Verzerrungen gefunden werden. Es lassen sich aber auch (abgesehen von Alter und Geschlecht) kaum Aussagen darüber treffen, wie sich die nicht teilnehmenden Lehrkräfte von den teilnehmenden unterscheiden.

Hinweise darauf, wieso die Mehrzahl der Lehrkräfte nicht teilnahm, geben die Begründungen der Schulleiter der 15 Schulen, die zum zweiten Befragungszeitpunkt nicht erneut teilnehmen wollten. Sie begründeten dies mit zahlreichen weiteren empirischen Befragungen, beispielsweise der Pilotierung von Bildungsstandards auf Bundes- und Landes-

Tab. 2: Alter und Geschlecht

	M Alter	Anteil weiblich	*N*
T1 (2005)			
Population[a]	47,7	56,8 %	13.510
Erste Querschnittstichprobe	46,3	57,6 %	914
T2 (2009)			
Population[a]	46,4	60,7 %	14.234
Zweite Querschnittstichprobe	45,7	63,1 %	734
T1 und T2			
Längsschnittstichprobe (Angaben zu T1)	46,0	56,0 %	141

[a]Realschullehrkräfte in Baden-Württemberg, elektronische Auskünfte des Statistischen Landesamtes Baden-Württemberg vom 28. Oktober 2010 und 11. November 2010

ebene, Befragungen anderer Hochschulen oder auch den Vorbereitungen zur externen Evaluation im Rahmen Neuer Steuerung. Weitere Gründe, die benannt wurden, waren „zu viel Arbeit“ oder das „fehlende Interesse der Lehrkräfte”. Diese Antworten deuten darauf hin, dass eher die Lehrkräfte an der Befragung teilnahmen, die selbst ein Interesse an der Einführung der Bildungsstandards und Vergleichsarbeiten hatten.

7.2 Instrumente

Der eingesetzte Fragebogen enthielt Fragen zur Einschätzung verschiedener Aspekte der Bildungsplanreform. Die (vermuteten) Auswirkungen der Vergleichsarbeiten wurden mit sieben Items erfragt (je drei zur Orientierungsfunktion und zu narrowing-the-curriuculum-Effekten, ein Item zu Übungserfordernissen), die auf der Grundlage qualitativer Interviews entwickelt und in zwei Pretests erprobt wurden. Die Lehrkräfte wurden gebeten einzuschätzen, wie sehr sie Aussagen zustimmen, die mögliche Auswirkungen der Vergleichsarbeiten beschreiben. Zur Beantwortung stand eine fünfstufige Likertskala (von *stimme nicht zu* bis *stimme zu*) zur Verfügung. Außerdem war es möglich anzugeben, dass das Item nicht eingeschätzt werden kann (*kann dazu nichts sagen*). Die Fragen zum ersten Messzeitpunkt (2005) waren in die Zukunft gerichtet, da die Vergleichsarbeiten zu diesem Befragungszeitpunkt noch nicht eingeführt waren. In der zweiten Befragung wurden die Items erneut vorgelegt, sie wurden aber nicht mehr auf die Zukunft, sondern auf die Gegenwart bezogen. Mit Faktorenanalysen (explorative Hauptkomponentenanalyse mit Varimax-Rotation) konnten die Items zur Orientierungsfunktion zu beiden Messzeitpunkten zu einem Faktor reduziert werden. Die narrowing-the-curriculum-Items ließen sich nicht konsistent auf einem Faktor abbilden. Sie werden deshalb nicht zu einer Skala zusammengefasst. Die sieben Items lauten im Einzelnen (Hier ist die Formulierung in der Gegenwart-Version angegeben – in der prospektiven Version lauten sie wie folgt: „Die künftigen Vergleichsarbeiten werden…“):

1. *Orientierungsfunktion* (Cronbachs $\alpha_{t1}=0{,}77$, $\alpha_{t2}=0{,}80$): drei Items, zur Unterrichtsplanung („Die Vergleichsarbeiten bieten mir einen Orientierungsrahmen zur Unterrichtsplanung“), Leistungsbeurteilung („Die Vergleichsarbeiten bieten mir einen Maßstab zur Leistungsbeurteilung“) und dem Erkennen von Lernrückständen („Durch die Vergleichsarbeiten kann ich die Lernrückstände einzelner Schüler erkennen und mir Maßnahmen zur Abhilfe überlegen“).
2. *narrowing-the-curriculum*: drei Items, zur Konzentration auf Kompetenzbereiche („Ich vermittle hauptsächlich jene Kompetenzbereiche, die in den Vergleichsarbeiten abgeprüft werden“), Aufgabenstellungen („Die Vergleichsarbeiten führen mich dazu, meinen Unterricht sehr stark an der Art ihrer Aufgabenstellungen zu orientieren“) und Fächer („Die Vergleichsarbeiten führen zur Vernachlässigung derjenigen Fächer, die nicht abgeprüft werden“).
3. erhöhte *Übungserfordernisse* („Weil sich die Vergleichsarbeiten auf Themen und Kompetenzen von zwei Schuljahren beziehen, muss ich mit meinen Klassen sehr viel mehr üben und wiederholen als bisher“).

7.3 Statistische Analysen

Wie sich die Einschätzungen vor und nach Einführung der Vergleichsarbeiten voneinander unterscheiden, wird zum einen anhand der zwei Querschnittstichproben (Personen, die entweder zum ersten oder zum zweiten Befragungszeitpunkt teilgenommen haben) geprüft und zum anderen anhand der Substichprobe der Personen, die zu beiden Messzeitpunkten teilgenommen haben (Längsschnittstichprobe). Die Querschnittstichproben bieten den Vorteil, einen größeren Teilnehmerkreis zu berücksichtigen. Die Längsschnittstichprobe erlaubt eine zusätzliche Absicherung der Ergebnisse und die Messung der Stärke des Zusammenhangs zwischen den Einschätzungen zu beiden Messzeitpunkten. Da für alle Personen in der Längsschnittstichprobe die Schulzugehörigkeit in Form einer Schul-ID bekannt ist, konnte die Mehrebenenstruktur der Daten berücksichtigt werden, indem für die Schulcluster adjustierte Standardfehler verwendet wurden, die mit Mplus (Version 6, Muthén und Muthén 1998–2010) berechnet worden sind.

Fehlende Werte sind überwiegend darauf zurückzuführen, dass Lehrkräfte angaben, Items nicht einschätzen zu können (*kann dazu nichts sagen*; zwischen 15,9 % und 27,7 %, je nach Item und Kohorte, $M = 21{,}7\,\%$, $SD = 3{,}6\,\%$). Keine Angaben machten nur vergleichsweise wenige Personen (zwischen 1,8 % und 3,3 %, $M = 2{,}5\,\%$, $SD = 0{,}5\,\%$). Die Skala *Orientierungsfunktion* wurde aus dem Mittelwert der zugehörigen Items gebildet, zu denen gültige Angaben vorlagen. Die Ergebnisse, die im folgenden Kapitel vorgestellt werden, beruhen ebenfalls auf denjenigen Personen, für die bei den jeweils benötigten Variablen gültige Angaben vorlagen (Fälle mit fehlenden Werten wurden exkludiert).

8 Ergebnisse

Querschnittstichproben. Die Stichprobe, die vor der Einführung befragt wurde (vgl. Tab. 3), erwartete insbesondere höhere Übungserfordernisse und Orientierungsleistungen durch die Einführung der Vergleichsarbeiten ($M > 3{,}70$). Dass narrowing-the-curriculum-Effekte auftreten, wurde moderat befürchtet ($3{,}00 < M < 3{,}50$).

In allen diesen Aspekten wurden die Vergleichsarbeiten vier Jahre später signifikant geringer bewertet als zum ersten Messzeitpunkt: Die Orientierungsleistung der Vergleichsarbeiten wird vor der Implementation und dem konkreten Umgang durch die Lehrkräfte

Tab. 3: T-Tests für unabhängige Stichproben: T1 vs. T2 (Querschnittvergleich)

	T1			T2			*T*	Cohens
	n	*M*	*SD*	*n*	*M*	*SD*		*D*
Orientierungsfunktion	652	3,71	1,06	435	2,42	1,17	18,43***	1,16
NTC-Kompetenzbereiche	632	3,36	1,32	426	2,30	1,24	13,02***	0,83
NTC-Aufgabenformate	623	3,42	1,28	428	2,03	1,15	18,39***	1,14
NTC-Fachvernachlässigung	592	3,16	1,50	395	2,23	1,44	9,75***	0,63
Übungserfordernisse	619	3,81	1,23	408	2,55	1,42	14,55***	0,95

NTC narrowing-the-curriculum

***$p < 0{,}001$

höher eingeschätzt als nach der vierjährigen Erfahrung mit ihnen. Der Unterschied in der Einschätzung ist substanziell und beträgt über eine Standardabweichung ($d = 1{,}16$). Auch der postulierte narrowing-the-curriculum-Effekt wird von den Lehrkräften zum zweiten Messzeitpunkt weit verhaltener eingeschätzt, dies gilt insbesondere für den Aspekt der Aufgabenformate. Ein substanzieller Rückgang in der Einschätzung ist ebenso in den Einzelitems zur Erfordernis von Wiederholungen und Übungen zu erkennen.

Längsschnittstichprobe. Die Daten der Längsschnittstichprobe vermögen diese Befunde zu replizieren. Die Bewertungen der Vergleichsarbeiten durch die Längsschnittstichprobe unterscheiden sich nicht signifikant von denjenigen der Querschnittstichprobe zu T1 (für die einzuschätzenden Aspekte: alle $|t| < 0{,}80$, alle $p > 0{,}40$) und nicht signifikant von denjenigen der Querschnittstichprobe zu T2 (alle $|t| < 1{,}50$, alle $p > 0{,}10$).

Die erwarteten Orientierungsleistungen und Wirkungen der Vergleichsarbeiten vor ihrer Einführung korrelieren mit den Einschätzungen vier Jahre später positiv: Für die Orientierungsfunktion beträgt die Korrelation $r = 0{,}25$, für die narrowing-the-curriculum-Items = 0,18 (Kompetenzbereiche), $r = 0{,}24$ (Aufgabenstellungen) bzw. $r = 0{,}20$ (Fachvernachlässigungen), für Übungserfordernisse $r = 0{,}31$. Mit Ausnahme der Items zur Konzentration auf bestimmte Kompetenzbereiche und Fachvernachlässigung sind die Korrelationen signifikant. In Anbetracht des Zeitabstandes beider Messungen von vier Jahren und der unterschiedlichen Perspektive bei der Einschätzung der Effekte (prospektiv und retrospektiv) sind diese Korrelationen als substanziell anzusehen: Wer einer Aussage vor der Einführung der Vergleichsarbeiten zugestimmt hat, tendiert auch dazu, diese zum zweiten Zeitpunkt entsprechend zu bewerten. Die Unterschiede in den Einschätzungen zu beiden Messzeitpunkten entsprechen im Großen und Ganzen denen, die bei den Querschnittstichproben gefunden wurden: In allen einzuschätzenden Aspekten ist die Zustimmung vier Jahre nach Einführung der Vergleichsarbeiten geringer als vor ihrer Einführung. Das gilt gleichermaßen für positive wie negative Aspekte (Tab. 4 und Abb. 1).

Tab. 4: Wald-Test für abhängige Stichproben: T1 vs. T2 (Längsschnittvergleich)

	T1		T2		*n*	T_W	Cohens
	M	*SD*	*M*	*SD*			*D*
Orientierungsfunktion	3,74	0,99	2,44	1,13	135	108,31***	1,05
NTC-Kompetenzbereiche	3,42	1,28	2,27	1,21	132	43,53***	0,84
NTC-Aufgabenformate	3,49	1,25	1,95	1,11	134	171,21***	1,11
NTC-Fachvernachlässigung	3,04	1,55	2,04	1,23	128	39,28***	0,66
Übungserfordernisse	3,77	1,22	2,46	1,29	132	109,43***	0,93

Die Wald-Tests wurden mit Mplus (Version 6; Muthén und Muthén 1998–2010) durchgeführt. Es wurden Standardfehler verwendet, die für die Klumpenstruktur der Daten (Schulzugehörigkeit zu T2 als Clustervariable) adjustiert wurden

***$p < 0{,}001$

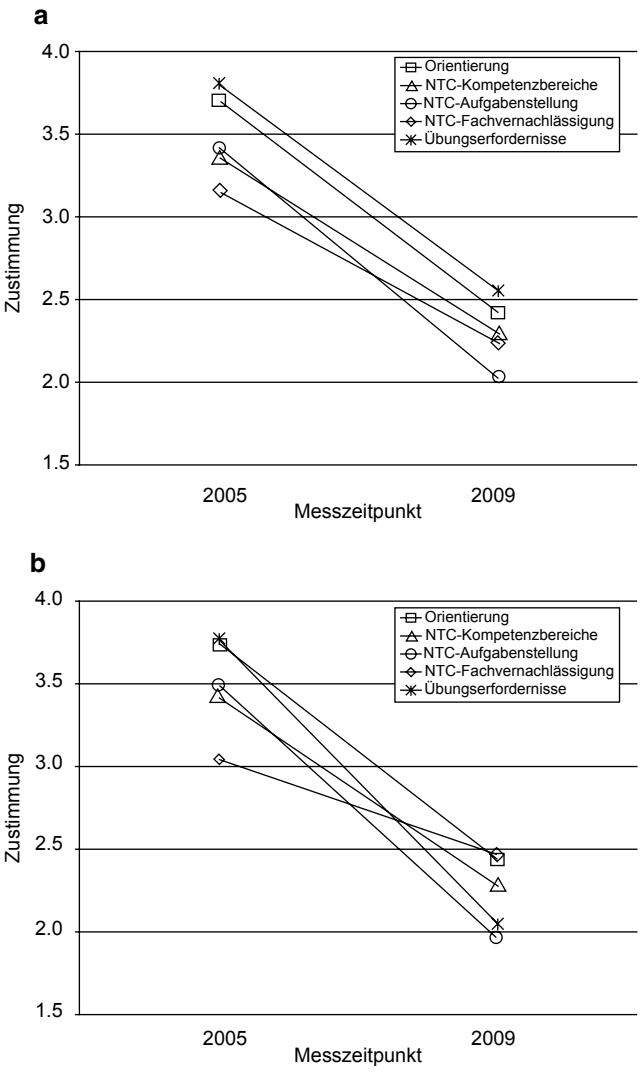


Abb. 1: Einschätzung der Vergleichsarbeiten vor (2005) und vier Jahre nach (2009) ihrer Einführung. **a** Querschnittstichproben. **b** Längsschnittstichprobe. *NTC* narrowing-the-curriculum; Zustimmung von 1 = *stimme nicht zu* bis 5 = *stimme zu*

9 Zusammenfassung und Ausblick

Im Mittelpunkt dieses Beitrags steht die Frage, wie Lehrkräfte (die zentralen Adressaten von Vergleichsarbeiten) der Einführung der Vergleichsarbeiten gegenüberstanden: Welche intendierten und nicht-intendierten Wirkungen erwarteten sie? Dazu wurden Realschullehrkräfte in Baden-Württemberg befragt und um ihre Einschätzung gebeten, ob die Vergleichsarbeiten ihrer Orientierungsfunktion zur Unterrichtsplanung, Leistungsbeurteilung und dem Erkennen von Lernrückständen gerecht werden und ob Befürchtungen gerechtfertigt sind, ihre Einführung ginge mit narrowing-the-curriculum-Effekten, der Vernachlässigung einzelner Fächer und vermehrten Übungserfordernissen einher. Diese Fragen wurden mit zwei querschnittlichen Datensätzen aus den Jahren 2005 (vor Einfüh-

rung der Vergleichsarbeiten) und 2009 (vier Jahre nach ihrer Einführung) zu beantworten versucht. Eine Teilstichprobe nahm zu beiden Befragungszeitpunkten an der Erhebung teil, sodass auch eine längsschnittliche Absicherung der Ergebnisse möglich war. Während zum ersten Messzeitpunkt die Lehrkräfte dabei ihre prospektiven Erwartungen an die Vergleichsarbeiten formulierten und noch keine Erfahrung mit der Durchführung der Vergleichsarbeiten haben konnten, waren nach vier realisierten Durchgängen die Vergleichsarbeiten in Baden-Württemberg etabliert.

Bezüglich der Orientierungsfunktion als intendierter Wirkung der Vergleichsarbeiten zeigte sich erwartungskonform ein Abfall in den Einschätzungen durch die Lehrkräfte: Während diese in der prospektiven Erwartung mehrheitlich zustimmten, dass ihnen die Vergleichsarbeiten einen Orientierungsrahmen zur Unterrichtsplanung, zur Leistungsbeurteilung und zum Erkennen von Lernrückständen der Schülerinnen und Schüler zu geben vermögen, war diese Auffassung nach vierjähriger Durchführung deutlich zurückgegangen und die Lehrkräfte antworteten im Durchschnitt auf der ablehnenden Skalenseite. Ein erheblicher Rückgang in der Bewertung konnte auch bezüglich der nicht intendierten Wirkungen festgestellt werden (narrowing-the-curriculum, höhere Übungserfordernisse). Vergleichsarbeiten an Realschulen in Baden-Württemberg wurden damit vier Jahre nach ihrer Einführung als weniger wirkungsvoll eingeschätzt als dies vor ihrer Einführung erwartet und – im Hinblick auf die nicht intendierten Wirkungen – befürchtet worden war.

Die Untersuchung vermag damit schon in vorhergehenden Studien gezogene Befunde auf der Basis einer schulformspezifisch breiteren Datengrundlage abzusichern und diese auf die konkrete Konzeption und Durchführung von Vergleichsarbeiten zu beziehen.[8] Sie ermöglicht in dieser Hinsicht eine über bisherige Studien hinausgehende längsschnittliche Betrachtung und ebenso einen in der Forschung angemahnten „Vorher-Nachher-Vergleich“, wie dieser bislang nicht aufzufinden ist. Darüber hinaus repräsentiert die gezogene Stichprobe über die Stratifizierungsmerkmale die Schulform Realschule in Baden-Württemberg adäquater als in zuvor erfolgten Untersuchungen.

Einschränkend ist jedoch anzumerken, dass lediglich die subjektiven Einstellungen der Lehrkräfte erhoben wurden. Weitere Forschungen, die die konkrete Erfahrung der Lehrkräfte im Umgang mit den Arbeiten auch fachspezifisch in den Blick zu nehmen vermögen, ist darauf aufbauend erforderlich. Die Befunde gelten ausschließlich für teilnehmende Realschulen im Bundesland Baden-Württemberg. Schlüsse auf ähnliche flächendeckende Assessments in weiteren Bundesländern oder Stadtstaaten sind nicht zulässig, da dort unterschiedliche Makrokonfigurationen der Steuerung vorliegen, die den Vergleichsarbeiten jeweils differente Funktionen in der Gesamtsteuerung zuweisen. Neben der eingeschränkten Generalisierbarkeit der Befunde auf andere Schulformen und Bundesländer ist bei der Interpretation der Daten vor allem die Rücklaufquote zu berücksichtigen, die in der vorliegenden Studie bei ca. 25 % lag, einem geringen – aber in Studien zu Vergleichsarbeiten üblichen – Wert. Es ist davon auszugehen, dass eher an den Bildungsstandards und Vergleichsarbeiten interessierte Lehrkräfte an der Befragung teilnahmen (vgl. Abschn. 7.1). Weniger daran interessierte Lehrkräfte werden vor Einführung der Vergleichsarbeiten vermutlich keine stärkeren Wirkungen von diesen erwarten als die eher interessierten. Die vorliegenden Daten erlauben jedoch keine Prüfung dieser Hypothese und keine Prüfung, in welchen Merkmalen sich die teilnehmenden Lehrkräfte von den nicht-teilnehmenden unterscheiden. Da jede Lehrkraft an der Realschule in

Baden-Württemberg ein Hauptfach studiert und in diesem durch die fachliche Gliederung der Realschulen eingesetzt wird, gehen wir davon aus, dass etwas über zwei Drittel bis drei Viertel der Lehrkräfte im Zeitraum zwischen dem ersten und zweiten Messzeitpunkt in die Durchführung der Vergleichsarbeiten involviert waren. Eine weitere Limitation der Untersuchung liegt darin, dass im Forschungsprojekt nicht erhoben werden konnte, welche Ergebnisse die antwortende Lehrkraft in den durchgeführten Vergleichsarbeiten erzielten und welche Wirkungen daraus auf die getroffenen Einschätzungen resultieren. Diese Informationen sind bedeutsam für die Analyse, weil die Resultate positive oder negative Auswirkungen auf die Beantwortung der Forschungsfragen nach sich ziehen. Die Beantwortung dieser wichtigen Fragen muss einem nachfolgenden Forschungsprojekt vorbehalten bleiben.

Welche möglichen Gründe können dafür angeführt werden, dass die Wirkung der Vergleichsarbeiten vier Jahre nach ihrer Implementation geringer eingeschätzt wurde als vor ihrer Implementation erwartet? Weil die Rückmeldeformate in Baden-Württemberg mit der Bezugnahme auf Leistungsgruppen ausschließlich die soziale Bezugsnorm zugrunde legen, bleibt unter Verweis auf den deutschen Forschungsstand zu vermuten, dass einerseits die Verwendung kriterialer Vergleiche in Form von Kompetenzstufen und weiterführenden Fehleranalysen sowie andererseits die Verwendung auf Referenzgruppen bezogener Landesmittelwerte (beispielsweise SES-korrigierter Landesmittelwerte) die wahrgenommene Orientierungsfunktion der Lehrkräfte erhöht. Auch eine verbesserte psychometrische Qualität der Vergleichsarbeiten unterstützt dieses Ziel. Mit Blick auf die Testerfahrungen der Lehrkräfte ergeben sich auch Hinweise darauf, dass externe Unterstützung bzw. Fortbildungen zur Analyse und Interpretation der Ergebnisse, welche auch die Differenz von Test- und Lernaufgaben beinhaltet, die wahrgenommene Orientierungsfunktion durch die Lehrerinnen und Lehrer verbessert.

Vergleichsarbeiten an Realschulen in Baden-württemberg scheinen deshalb aus Sicht der Lehrkräfte „weder nützlich noch schädlich“ zu sein. Bis zum Zeitpunkt der Befragung 2009 sehen die Lehrkräfte keine maßgebenden Impulse für die Unterrichtsentwicklung, wie dies eigentlich intendiert war. Die Befunde legen im Vergleich der intendierten mit den realisierten Funktionen der Tests den Schluss nahe, die Konzeption und Durchführung der Vergleichsarbeiten und ihre Stellung innerhalb der Gesamtkonfiguration der Steuerung neu zu prüfen und zu überdenken.

Danksagung: Die Autoren danken den Gutachtern der ZfE, Herrn Prof. Dr. Drs. h.c. Jürgen Baumert sowie Christiane Fiege für wertvolle Hinweise zu diesem Aufsatz.

Anmerkungen

1 Auskunft des Landesinstituts für Schulentwicklung in Stuttgart vom 29. Mai 2012.

2 Literatur zu diesem Bereich liegt sehr eingeschränkt vor. Die Ausführungen beziehen sich deshalb auf Informationen aus Gesprächen, die die Verfasser mit Lehrkräften von Realschulen geführt haben.

3 Auskunft des Landesinstituts für Schulentwicklung in Stuttgart vom 29. Mai 2012.

4 Auskunft des Landesinstituts für Schulentwicklung in Stuttgart vom 29. Mai 2012.

5 Zum Schuljahr 2010/2011 wurde in Baden-Württemberg diese Praxis wieder geändert, ihre Bewertung fließt nun nicht mehr in die schulische Leistungsbeurteilung ein.

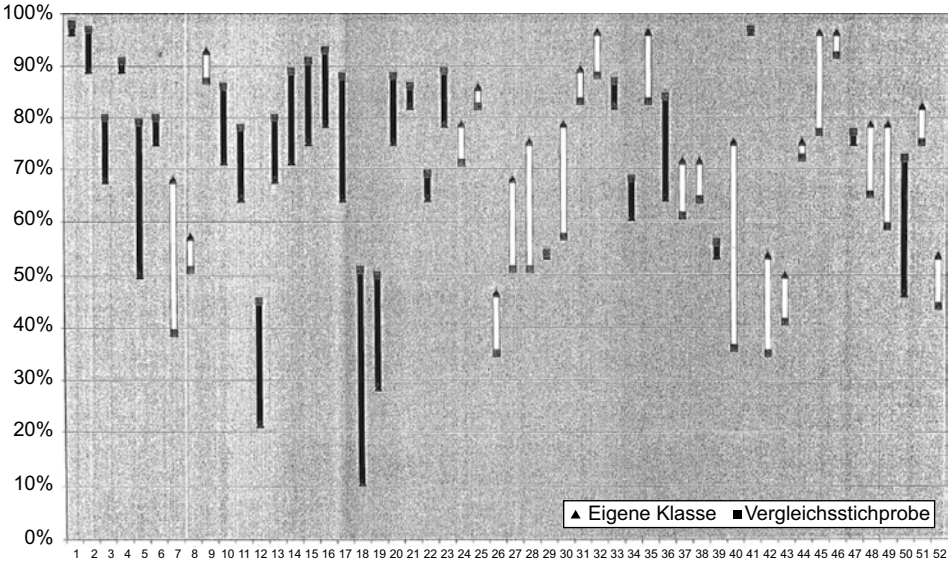
6 Die Befragung wurde von Maier (2008b) auch auf das Bundesland Thüringen ausgeweitet. Die Vergleichsarbeiten wurden von 311 Lehrkräften in Thüringen sämtlich besser bewertet als von 825 Lehrkräften, die aus Baden-Württemberg teilnahmen. Lediglich die Orientierungsleistung der Vergleichsarbeiten für die Notengebung wird von den Lehrkräften in Baden-Württemberg besser eingeschätzt. Diesen Befund führte Maier auf eine Differenz in den Rückmeldeformaten zurück. Hier erfolgten in Thüringen im Unterschied zu Baden-Württemberg Rückmeldungen, die SES-korrigierte Landesmittelwerte ebenso enthielten wie eine Aufgliederung der Klassenwerte in fachspezifische Kompetenzen und weiterführenden Fehleranalysen. Nach Maier war die Informiertheit der Lehrkräfte zu Beginn der Reform vergleichsweise niedrig und stieg danach an (Maier 2009). Zusammenfassend weisen die Ergebnisse einerseits auf eine Orientierungsleistung, die von den flächendeckenden Assessments für die Lehrkräfte ausgeht, hin, andererseits darauf, dass Wirksamkeit und Akzeptanz von Vergleichsarbeiten im Zeitverlauf abnahmen. Die Befürchtung von curricularen Verengungen wird von den Lehrkräften nicht geteilt und insgesamt keine „extreme Deformation des Unterrichts“ (Maier 2010) festgestellt. Auch Hinweise darauf, dass schulische Bildungsziele eingeschränkt werden, konnten nicht aufgefunden werden (Maier 2009).

7 Elektronische Auskunft des Statistischen Landesamtes vom 11. November 2010.

8 Häufig wird in Untersuchungen die konkrete Durchführungspraxis, die bedeutsam für die Interpretation der Befunde ist, nur unzureichend beschrieben.

Anhang A

Beispiel für eine Ergebnisrückmeldung auf Aufgabenebene, 6. Klasse Deutsch. (Die Rückmeldung wurde vom Landesinstitut für Schulentwicklung 2006 erstellt)



Anhang B

Beispiel für eine Ergebnisrückmeldung auf Klassenebene, 7. Klasse Mathematik (Die Rückmeldung wurde vom Landesinstitut für Schulentwicklung 2006 erstellt.)

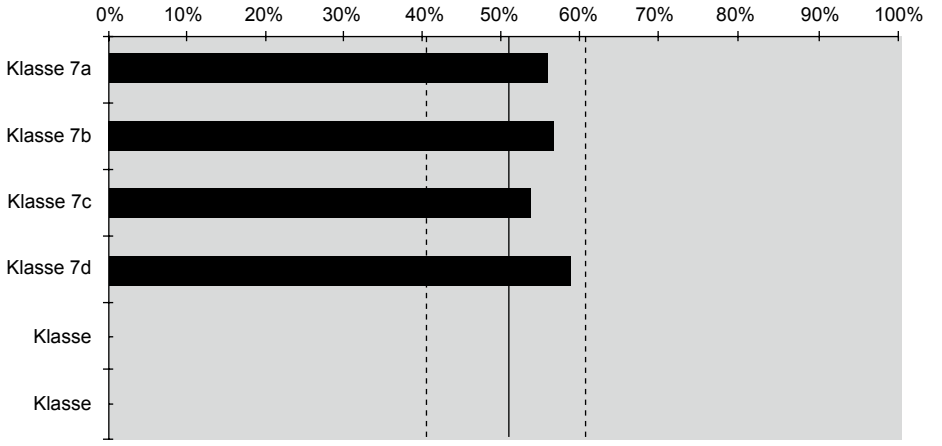
Detenauswertung - Gesamttest
Mathematik Klassen 7

Wie viele Items des Gesamttestes wurden im Mittel von den Schülerinnen und Schülern der jeweiligen Klassen (Kurse, Gruppen) gelöst?

Anzahl aller Items: 35

Klasse / Anzahl Schüler/innen die teilgenommen haben		Anzahl durchschnittlich gelöster Artikel in der Klasse	Anzahl durchschnittlich gelöster Artikel in der Klasse (in Prozent)
Klasse 7a	30	20	56,0%
Klasse 7b	28	20	57,0%
Klasse 7c	30	19	54,0%
Klasse 7d	26	21	59,0%
Klasse	0	0	0,0%
Klasse	0	0	0,0%
Vergleichsstichprobe		18	51,0%

Inwieweit weichen die jeweiligen Klassenergebnisse von der Vergleichsstichprobe ab?
Die durchgezogene rote Referenzlinie kennzeichnet den landesweiten Vergleichswert, die gestrichelten Linien stellen die Differenz von +/- 10% dar. Diese und darüber hinausgehende Abweichungen werden als bedeutsam betrachtet. Dies gilt auch für Unterschiede zwischen den einzelnen Klassen.



Literatur

Beer, R. (2006). Qualitätsentwicklung durch Bildungsstandards? Ergebnisse einer Befragung der betroffenen Lehrerinnen und Lehrer in Wien – 2005. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 253–264). Münster: Waxmann.

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

Berkemeyer, N., & Holt, N. van (2012). Leistungsrückmeldungen im Längsschnitt – Erste Erfahrungen mit dem Schüler-Monitoring-System (SMS). In A. Wacker, U. Maier, & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung. Empirische Befunde und forschungsmethodische Implikationen* (S. 109–130). Wiesbaden: VS Verlag für Sozialwissenschaften.

Böttcher, W., & Dicke, J. N. (2008). Implementation von Standards. Empirische Ergebnisse einer Umfrage bei Deutschlehrern. In W. Böttcher, W. Bos, H. Döbert, & H. G. Holtappels (Hrsg.), *Bildungsmonitoring und Bildungscontrolling in nationaler und internationaler Perspektive. Dokumentation zur Herbsttagung der Kommission Bildungsorganisation, – planung, – recht (KBBB)* (S. 143–156). Münster: Waxmann.

Bonsen, M., & von der Gathen, J. (2004). Schulentwicklung und Testdaten. Die innerschulische Verarbeitung von Leistungsrückmeldungen. In H. G. Holtappels, K. Klemm, H. Pfeiffer, H.-G. Rolff, & R. Schulz-Zander (Hrsg.), *Jahrbuch für Schulentwicklung 13* (S. 225–252). Weinheim: Juventa.

Carlgren, I. (2009). The Swedish comprehensive school – lost in transition? *Zeitschrift für Erziehungswissenschaft, 12,* 633–649.

Chorny, V., & Webbink, D. (2010). *The effect of accountability policies in primary education in Amsterdam. CPB Discussion Paper.* Amsterdam: Centraal Planbureau.

Drieschner, E. (2009). *Bildungsstandards praktisch. Perspektiven kompetenzorientierten Lehrens und Lernens.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Erickson, G., & Lander, R. (2007). Der Kitt, der ein wachsendes System zusammen hält? Nationale Tests als Kern der Qualitätssicherung in Schweden. *Pädagogik, 59*(3), 32–35.

Fiege, C., Reuther, F., & Nachtigall, C. (2011). Faire Vergleiche? – Berücksichtigung von Kontextbedingungen des Lernens beim Vergleich von Testergebnissen aus deutschen Vergleichsarbeiten. *Zeitschrift für Bildungsforschung, 1,* 133–149.

Gördel, B. (2008). Die Einführung der nationalen Bildungsstandards in drei Bundesländern – eine explorative Studie zu Implementierungsstrategien. In R. Langer (Hrsg.), *„Warum tun die das?" Governanceanalysen zum Steuerungshandeln in der Schulentwicklung* (S. 193–220). Wiesbaden: VS Verlag für Sozialwissenschaften.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., et al. (2007). *Standards-based accountability under no child left behind: Experiences of teachers and administrators in three states*. Santa Monica: RAND Corporation.

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

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

Herman, J. L. (2004). The effects of testing on instruction. In S. H. Fuhrmann & R. F. Elmore (Hrsg.), *Redesigning accountability systems for education* (S. 141–166). New York: Teachers College Press.

Herrmann, U. (2005). Fördern ‚Bildungsstandards' die allgemeine Schulbildung? In J. Rekus (Hrsg.), *Bildungsstandards, Kerncurricula und die Aufgabe der Schule. Münstersche Gespräche zur Pädagogik* (S. 24–52). Münster: Aschendorff.

Herrmann, U. (2010). *Schulen zukunftsfähig machen*. Bad Heilbrunn: Klinkhardt.

Hursh, D. (2005). The growth of high-stakes testing in the USA: Accountability, markets and the decline in educational equality. *British Educational Research Journal, 31,* 605–622.

Innenministerium Baden-Württemberg (Hrsg.). (1984). *Landesentwicklungsplan Baden-Württemberg vom 12. Dezember 1983 mit Begründung und Anlagen*. Freudenstadt: VUD-Verlag.

Koretz, D. M., Barron, S., Mitchell, K. J., & Stecher, B. M. (1996a). *Perceived effects of the Kentucky instruction results information district*. Santa Monica, CA: RAND Corporation.

Koretz, D. M., Mitchell, K. J., Barron, S. I., & Keith, S. (1996b). *The perceived effects of the Maryland School Performance Assessment Program* (*CSE Technical Report No. 409).* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Künzli, R. (2006). Standards statt Lehrpläne – zurück zu den Bildungsinhalten? In L. Criblez, P. Gautschi, P. H. Monico, & H. Messner (Hrsg.), *Lehrpläne und Bildungsstandards. Was Schülerinnen und Schüler lernen sollen. Festschrift zum 65. Geburtstag von Prof. Dr. Rudolf Künzli* (S. 83–102). Bern: h.e.p-verlag.

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

Landesbildungsserver Baden-Württemberg. (2009). Vergleichsarbeiten „DVA". Konzeption-Ziele. Stuttgart: Ministerium für Kultus, Jugend und Sport. http://www.schule-bw.de/entwicklung/dva/vadva/ konzeption-dva/ziele/. Zugegriffen: 9. Mai 2011.

Landesinstitut für Schulentwicklung Baden-Württemberg. (2010). Vergleichsarbeiten „DVA". Umgang mit den Ergebnissen im Rahmen der Selbstevaluation der Schulen. http://www.schule-bw.de/entwicklung/dva/dva_docs/docs4dva/Handreichung_Umgang.pdf. Zugegriffen: 5. Juli 2012.

Lipman, P. (2004). *High-stakes education: Inequality, globalization, and urban school reform.* New York: Routledge Falmer.

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 Schulssytem* (S. 145–169). Wiesbaden: Verlag für Sozialwissenschaften.

Maier, U. (2008a). Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. *Zeitschrift für Pädagogik, 54,* 95–115.

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

Maier, U. (2009). *Wie gehen Lehrerinnen und Lehrer mit Vergleichsarbeiten um? Eine Studie zu testbasierten Schulreformen in Baden-Württemberg und Thüringen*. Baltmannsweiler: Schneider Hohengehren.

Maier, U. (2010). Effekte testbasierter Rechenschaftslegung auf Schule und Unterricht. Ist die internationale Befundlage auf Vergleichsarbeiten im deutschsprachigen Raum übertragbar? *Zeitschrift für Pädagogik, 56,* 112–128.

Maier, U., & Rauin, U. (2006). Vergleichsarbeiten – Hilfe zur Unterrichtsentwicklung? Zentrale Lernstandserhebungen aus Sicht baden-württembergischer Lehrkräfte. *Die Deutsche Schule, 98,* 403–435.

McDonnell, L. M., & Choicer, C. (1997). *Testing and teaching: Local implementation of new state assessments (CSE Technical Report No. 442).* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus – Statistical analysis with latent variables.* Los Angeles: Muthén & Muthén.

Nachtigall, C., & Jantowski, A. (2007). Die Thüringer Kompetenztests unter besonderer Berücksichtigung der Evaluationsergebnisse zum Rezeptionsverhalten. *Empirische Pädagogik, 21,* 401–410.

National Commission on Excellence in Education. (Hrsg.). (1983). *A nation at risk: The imperative for educational reform*. Washington: U.S. Government Printing Office.

Oelkers, J., & Reusser, K. (2008). *Qualität entwickeln – Standards sichern – mit Differenz umgehen. Eine Expertise.* Berlin: Bundesministerium für Bildung und Forschung.

Plöger, W. (2005). Aus der Vergangenheit lernen? – Bildungsstandards unter historisch-systematischer Perspektive. In J. Rekus (Hrsg.), *Bildungsstandards, Kerncurricula und die Aufgabe der Schule. Münstersche Gespräche zur Pädagogik* (S. 91–107). Münster: Aschendorff.

Regenbrecht, A. (2005). Sichern Bildungsstandards die Bildungsaufgabe der Schule? In J. Rekus (Hrsg.), *Bildungsstandards, Kerncurricula und die Aufgabe der Schule. Münstersche Gespräche zur Pädagogik* (S. 53–76). Münster: Aschendorff.

Rothe, C. (2003). Summative Leistungsmessung. Als Element der Qualitätssicherung und Bildungsreform an Realschulen in Baden-Württemberg. *PÄDForum, 31*(5), 262–265.

Schirp, H. (2006). Zentrale quantitative Leistungsmessungen und qualitative Schulentwicklung. Die Wirkungen von High Stakes Tests in den USA. *Die Deutsche Schule, 98,* 422–435.

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

Scholl, D. (2009). *Sind die traditionellen Lehrpläne überflüssig? Zur lehrplantheoretischen Problematik von Bildungsstandards und Kernlehrplänen.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Sikorski, P. B. (2003). Diagnose- und Jahrgangsarbeiten. Zur Entwicklung am Landesinstitut für Erziehung und Unterricht in Stuttgart. *PÄDForum, 31*(5), 260–261.

Stecher, B. M. (2002). Consequences of large-scale, high stakes testing on school and classroom practice. In L. S. Hamilton, B. M. Stecher, & S. P. Klein (Hrsg.), *Making sense of test-based accountability in education* (S. 79–100). Santa Monica, CA: RAND Corporation.

Stecher, B. M., & Barron, S. (2001). Unintended consequences of test-based accountability when testing in „milepost" grades. *Educational Assessment, 7,* 259–281.

Vollstädt, W., Tillmann, K.-J., Rauin, U., Höhmann, K., & Tebrügge, A. (1999). *Lehrpläne im Schulalltag. Eine empirische Studie zur Akzeptanz und Wirkung von Lehrplänen in der Sekundarstufe I.* Opladen: Leske + Budrich.

Wacker, A. (2008). *Bildungsstandards als Steuerungsinstrumente der Bildungsplanung. Eine empirische Studie zur Realschule in Baden-Württemberg*. Bad Heilbrunn: Klinkhardt.

Waldow, F. (2005). Späte Sanktionierung – Formen und Funktionen der Leistungsmessung in Schweden. In H. Döbert & H.-W. Fuchs (Hrsg.), *Leistungsmessungen und Innovationsstrategien in Schulsystemen. Ein internationaler Vergleich* (S. 125–138). Münster: Waxmann.