ZfE 11 (2008) 3:453-474
DOI 10.1007/s11618-008-0036-0

ALLGEMEINER TEIL

ZfE

Vergleichsarbeiten im Vergleich – Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessungen in Baden-Württemberg und Thüringen

Uwe Maier

Zusammenfassung: Testbasierte Schul- und Unterrichtsreformen sind mittlerweile ein beliebtes Instrument bundesdeutscher Bildungspolitik. Die Leistungsmessungen und die Art der Implementation unterscheiden sich jedoch von Bundesland zu Bundesland zum Teil erheblich. Besonders deutliche Differenzen sind zwischen den Kompetenztests in Thüringen (Klasse 6 und 8) und den Vergleichsarbeiten in Baden-Württemberg (Klasse 6) erkennbar. Beispielsweise sind die Thüringer Kompetenztests in ein umfangreiches Fortbildungs- und Weiterentwicklungskonzept eingebettet und die Lehrer erhalten kriteriale und faire Vergleiche als Rückmeldung. In einer quantitativen Lehrerbefragung in beiden Bundesländern (n=1136) wurde deshalb geprüft, ob diese Differenz zu unterschiedlichen Akzeptanz- und Nutzungseinschätzungen bei Lehrkräften führt und die schulinterne Diskussion über Testrückmeldungen beeinflusst. Es zeigte sich, dass Thüringer Lehrkräfte sowohl die Akzeptanz, die curriculare Validität als auch die Nutzung der Leistungsdaten für Lerndiagnosen und die zukünftige Unterrichtsplanung höher einschätzen. Lediglich die Nutzung der zentralen Leistungsrückmeldungen für die Notengebung wurde von baden-württembergischen Lehrkräften höher bewertet. In Thüringen sind die Leistungsrückmeldungen auch wesentlich häufiger Gegenstand systematischer Diskussionen in Fachkonferenzen.

Schlüsselwörter: Vergleichsarbeiten · Rückmeldungen · testbasierte Rechenschaftslegung · datenbasierte Schulentwicklung · Lehrer · Sekundarstufe · Bildungsstandards

Comparative Tests in Comparison – Acceptance and perceived use of standardized student assessment in Baden-Württemberg and Thuringia

Abstract: Test-based reforms of schools and teaching have become popular instruments of German educational policy. However, performance assessment and its implementation differ widely between the German states. The differences between the comparative tests used in Thuringia (years 6 and 8) and Baden-Württemberg (year 6) are particularly clear. For instance, the Thuringia competency tests are embedded within an extensive concept for teachers' continual professional development and the teachers receive fair, criteria-based feedback. A quantitative teacher survey in both German states (n=1136) was carried out in order to find out whether these differences led to differences in the respective teachers' acceptance and appraisal of the tests and whether the

Dr. Uwe Maier (✉)
Pädagogische Hochschule Schwäbisch Gmünd, Institut für Erziehungswissenschaft, Oberbettringerstraße 200, D-73525 Schwäbisch Gmünd, Deutschland
E-Mail: uwe.maier@ph-gmuend.de

test results initiated internal discussions amongst teachers. The results show that teachers from Thuringia had a higher acceptance and estimation of curricular validity, uses of the data for learning diagnosis and for future lesson planning. Singularly, the use of the central test results to mark students received a higher evaluation from teachers in Baden-Württemberg. The tests results are also more frequently the subject of systematic discussions in school forums in Thuringia.

Keywords: comparative tests · data-based school reform · educational standards · feedback · secondary school · teachers · test-based accountability

1. Problemstellung

Bereits wenige Jahre nach den KMK-Beschlüssen von 2003 zur Qualitätssicherung im Bildungssystem sind in allen Bundesländern zentrale Vergleichsarbeiten eingeführt worden (Überblick bei Sill/Sikora 2007). Stand anfänglich lediglich die Entwicklung von Testaufgaben und Kompetenzmodellen im Vordergrund, kann man mittlerweile eine sich verstärkende Diskussion über die Akzeptanz und Nutzung der neuen Instrumente wahrnehmen (z. B. bei den Vorträgen der 7. Tagung „Empiriegestützte Schulentwicklung“ in Mainz). Dabei stellt sich vor allem die Frage, ob durch zentrale Leistungsmessungen eine datenbasierte Schul- und Unterrichtsentwicklung unterstützt werden kann. Während Testprotagonisten sehr optimistisch davon ausgehen, dass Vergleichsarbeiten zumindest über einen längeren Zeitraum gesehen eine Wirkung entfalten können (z. B. Hosenfeld/Schrader/Helmke 2006; Tresch 2007), gibt es auch wesentlich pessimistischere Einschätzungen. Beispielsweise sehen Altrichter und Heinrich (2006) in zentralen Tests einen Bruch mit unausgesprochenen professionellen Prinzipien an Schulen und unterstellen der Bildungspolitik ein unterkomplexes Steuerungsmodell, das die institutionellen Unwägbarkeiten einer Implementation zentraler Tests nicht erkennt.

Unabhängig von dieser zum Teil normativ aufgeladenen Debatte über neue Steuerungsstrukturen im Bildungswesen findet man auf unterschiedlichen Theorie- und Abstraktionsebenen Vorschläge, wie zentrale Vergleichsarbeiten und Testrückmeldungen zu gestalten sind, um Lehrkräfte und Schulen über die Reflexion auch zu einer Weiterentwicklung und Verbesserung der eigenen Praxis anregen zu können. Prozess- oder Komponentenmodelle zur Lokalisierung der Bedingungen und Einflussfaktoren auf die schulinterne Nutzung von Leistungsdaten wurden mit unterschiedlicher theoretischer und empirischer Breite beispielsweise von Visscher und Coe (2003), Helmke und Hosenfeld (2005) oder Tresch (2007) vorgelegt. Auch in der internationalen Literatur zu „washback“-Effekten findet man theoretische Modelle, mit denen Möglichkeiten der Veränderung durch zentrale Tests beschrieben werden (Bailey 1996; Cheng/Curtis 2004). Dabei wird ebenfalls argumentiert, dass die allgemeine Schulentwicklungsforschung (school improvement research) bereits zentrale Bedingungen beschrieben hat, die auch auf die Wirksamkeit einer datengestützten Schulentwicklung übertragen werden können (Wall 2000; Hulpia/Valcke 2004).

Da sich die Modelle in der Regel stark überlappen und ähnliche Bedingungsfaktoren datenbasierter Unterrichtsentwicklung beschreiben, soll exemplarisch das Rahmenmo-

dell für „school performance feedback systems“ (SPFS) von Visscher und Coe (2003) skizziert werden. Die Autoren unterscheiden grundsätzlich den Entwicklungsprozess, die zentralen Merkmale des Testsystems, die Bedingungen der Implementation und relevante Schulmerkmale für die sinnvolle Nutzung von Leistungsdaten. Merkmale von Testsystemen, die die Nutzung beeinflussen, sind vor allem die Validität (value-added), die Reliabilität, die Aktualität und die Relevanz der Daten für die Lehrkräfte. Als förderliche Merkmale der Implementation werden die Einbettung der Tests in langfristig angelegte Reformstrategien, die Unterstützung im Umgang mit den Daten und eine Verknüpfung mit maßvollem Druck von außen angesehen. Ebenfalls entscheidend ist, ob Lehrkräfte im Zuge der Einführung des Testsystems ermutigt werden, die Daten selbstständig zu interpretieren (ownership of data). Bei der Datennutzung stellt sich ebenfalls die Frage, ob zusätzliche Ressourcen bereitstehen, um Innovationen voranbringen zu können. Wenn durch Leistungstests Bedarfe offenkundig werden, macht diese Information nur Sinn, wenn sie auch zur zielgerichteten Allokation vorhandener Ressourcen genutzt wird.

Auch lernpsychologische Modelle zu Feedbackeffekten eignen sich, um herauszufinden, welche Merkmale einer Leistungsrückmeldung für die produktive Weiterarbeit relevant sein könnten. Aus der „feedback intervention theory“ (Kluger/De Nisi 1996) oder dem „model of effective feedback“ (Hattie/Timperley 2007) lassen sich folgende Prämissen für Leistungsrückmeldungen ableiten: Eine Rückmeldung, die vom Lernenden (in diesem Fall dem Lehrer) als Lob aufgefasst wird, kann die Aufmerksamkeit auf die eigene Selbstwirksamkeit lenken und stellt somit keine verwertbaren Informationen für den weiteren Lernprozess zur Verfügung. Die Feedbackeffekte sind stärker, wenn sich die Hinweise auf die erbrachte Leistung oder die Lernprozesse beziehen und damit internale Attribuierungen stützen. Es stellt sich somit die Frage, ob eine Lehrkraft durch die zentrale Testrückmeldung neue, konkrete Hinweise erhält oder ob das Feedback als pauschale Kritik oder pauschales Lob verstanden wird.

Ohne die internationale Diskussion weiter zu vertiefen, soll nun auf konzeptionelle Überlegungen im deutschsprachigen Raum eingegangen werden. Vor allem im Rahmen der größeren Vergleichsarbeitsprojekte in Deutschland (Kompetenztests Thüringen, Lernstand NRW, VERA) wurde, vorwiegend aus einer mathematikdidaktischen Perspektive, der Frage nachgegangen, wie zentrale Tests sowohl messtheoretischen Anforderungen genügen, als auch für Lehrkräfte ein gewisses Innovationspotenzial entfalten können (Büchter/Leuders 2005; Blum et al. 2005; Peek et al. 2006; Lorenz 2005; Nachtigall/Kröhne 2006; Peek/Dobbelstein 2006; Peek/Steffens/Köller 2006; Sill/Sikora 2007). Diese konzeptionellen Überlegungen decken sich zum großen Teil mit den bereits genannten Modellen. Allerdings wird in diesen Texten die eingangs formulierte Frage nach der Möglichkeit einer datenbasierten Unterrichtsentwicklung noch einmal fachdidaktisch zugespitzt. Die wichtigsten Forderungen lassen sich folgendermaßen zusammenfassen:

- Als Standard gilt mittlerweile die kriteriale Rückmeldung von Kompetenzprofilen und der Aufgabenschwierigkeit auf Klassenebene. Diese Daten sind aufgrund des Aggregationsniveaus hinreichend reliabel und ermöglichen der Lehrkraft eine Gesamtbeurteilung der durch den Unterricht aufgebauten Kompetenzen.

- Korrigierte Vergleichswerte oder die Rückmeldung adjustierter Daten (value-added) unter Angabe von Konfidenzintervallen ermöglichen „faire“ Vergleiche mit Parallelklassen oder anderen Schulen.
- Zentrale Tests können vor allem dann hilfreich sein, wenn sie den Lehrkräften im Sinne von „opportunity-to-learn standards“ genug fachdidaktische Orientierung für den Unterricht geben. Die Schwierigkeit liegt vor allem darin, dass Testaufgaben einerseits psychometrischen und andererseits fachdidaktischen Anforderungen genügen müssen und darüber hinaus noch Impulse für die Schulentwicklung geben sollen.
- Die Lehrkräfte sollten ebenfalls bei einer gezielten Fehleranalyse unterstützt werden. Bei Multiple-Choice-Items ist dies in der Regel nur schwer möglich. Der Lösungsraum ist eingeschränkt und die Ratewahrscheinlichkeit lässt auch bei vier Distraktoren kaum zuverlässige Rückschlüsse auf Denk- und Lösungswege der Schüler zu.

Die theoretischen und konzeptionellen Überlegungen zu zentralen Tests und Rückmeldeformaten sind weitgehend konsistent und plausibel. Die Rückmeldung muss verständlich, praktikabel und für die Lehrkraft relevant sein (Peek et al. 2006). Nun stellt sich die Frage nach empirischen Evidenzen. In Deutschland wurde der Frage nach der Rezeption und Nutzung von Testergebnissen bereits im Rahmen von TIMSS, PISA und anderen „large scale assessments“ nachgegangen (z. B. Kohler 2004; Peek 2004; Klug/Reh 2000; Schrader/Helmke 2004). Dabei zeigte sich durchweg, dass die Rückmeldung hoch aggregierter Daten aus Stichprobenstudien von Lehrkräften nicht mit dem eigenen Unterricht in Verbindung gebracht wird.

Mittlerweile werden im Rahmen der aktuellen Vergleichsarbeitsprojekte regelmäßige Rezeptionsstudien durchgeführt (Nachtigall 2005; Groß Ophoff et al. 2006; Bonsen/Büchter/Peek 2006). Auch aus der Schweiz sind bereits entsprechende Untersuchungsergebnisse bekannt (Moser 2003; Keller/Moser 2006; Tresch 2007; Baeriswyl et al. 2006). Da die Vergleichsarbeitsprojekte und die Rezeptionsstudien zum Teil sehr unterschiedlich angelegt sind, lässt sich der aktuelle und noch sehr spärliche Forschungsstand nur schwer vergleichen. Allerdings sind bereits erste Tendenzen erkennbar.

Zentrale Leistungsmessungen werden von Lehrkräften nicht per se abgelehnt. Besonders deutlich wird dies bei einigen Schweizer Projekten, die eine freiwillige Teilnahme der Lehrkräfte vorsehen (Moser 2003; Tresch 2007). Die Rückmeldungen werden in der Regel als informativ und verständlich wahrgenommen. Bestimmte statistische Begriffe und grafische Darstellungsformen mussten zwar reduziert und vereinfacht werden. Insgesamt gelang es aber den meisten Testprojekten, eine lesbare und überschaubare Menge an Informationen zurückzuspielen.

Problematisch wird es bei der Frage der Interpretation. Die VERA-Projektgruppe konnte beispielsweise mit ihrer Begleiterhebung darauf aufmerksam machen, dass ein überwiegender Teil der Lehrkräfte erwartungswidrige Testergebnisse external attribuiert (Groß Ophoff et al. 2006). Ebenfalls sehr unzureichend und teilweise auch widersprüchlich sind die empirischen Befunde hinsichtlich der Nutzung der Testrückmeldungen. In der Online-Befragung zu VERA werden Unterrichtsentwicklungsmaßnahmen angegeben. Überzeugend sind auch die Evaluationsergebnisse zu Check 5 im Kanton Aargau (Tresch 2007). Die Lehrkräfte wurden dazu angehalten, ihre Testinterpretation und ihre abgeleiteten Maßnahmen schriftlich zu fixieren. Dies führte zu einer sehr hohen Quote

an berichteten Folgemaßnahmen. Allerdings basieren sämtliche Studien auf Selbstauskünften ohne Validierung durch eine unabhängige Perspektive.

Als sehr bedeutsam kristallisiert sich die Bedeutung des subjektiv empfundenen Nutzens zentraler Testrückmeldungen heraus. In einer Online-Evaluation der Lernstandserhebungen in NRW fanden Bonsen, Büchter und Peek (2006), dass die Verständlichkeit nur über den subjektiv eingeschätzten Nutzen auf mögliche Folgemaßnahmen wirkt. Die Verständlichkeit der Testrückmeldungen allein bewirkt keine Veränderung.

Weder konzeptionell noch empirisch geklärt sind die Bereiche professionellen Handelns, die durch Testrückmeldungen beeinflusst werden sollen bzw. verändert werden können. Dies hängt natürlich von der Zielsetzung des jeweiligen Systems ab. Während in NRW eine Nutzung der Daten für individuelle Förder- oder Selektionsdiagnosen bewusst ausgeschlossen wird, findet man beispielsweise in der Schweiz das Deutschfreiburger Übergangsmodell, bei dem ein zentraler Test die Selektionsdiagnose am Ende der Primarstufe unterstützen soll. Am allerwenigsten geklärt ist die systematische Einbindung der Rückmeldedaten in die interne Schulevaluation. Für Nachtigall (2005) beispielsweise besteht ein noch ungeklärtes Verhältnis von externer und interner Evaluation und der Rolle von zentralen Tests in diesem Geflecht.

2. Fragestellung

Es gibt eine Fülle theoretisch begründbarer und fachdidaktisch plausibler Vorschläge, wie Vergleichsarbeiten und Leistungsrückmeldungen zu gestalten sind, um die Akzeptanz an Schulen zu erhöhen und Lehrkräften die professionelle Nutzung der Daten zu erleichtern. Die Ergebnisse der ersten Rezeptionsstudien geben zudem Hinweise auf den schulinternen Umgang mit Daten, wobei durch die jeweils spezifischen Untersuchungsanlagen die Ergebnisse kaum vergleichbar sind. Allerdings gibt es bisher noch keine Untersuchungen, in denen die Auswirkungen verschiedener Vergleichsarbeiten gegenübergestellt werden. In dieser Studie wurde deshalb die dem Bildungsföderalismus geschuldete Situation genutzt, dass in den einzelnen Bundesländern zum Teil sehr unterschiedliche Testsysteme und Rückmeldeformate entwickelt wurden. Als besonders kontrastreich in dieser Hinsicht können die zentralen Leistungsmessungen in Baden-Württemberg und Thüringen angesehen werden. In beiden Bundesländern wurden im Laufe der letzten Jahre Vergleichsarbeiten bzw. Kompetenztests für das allgemeinbildende Schulwesen eingeführt. Getestete Fächer, Jahrgangsstufen und die Testhäufigkeit sind weitgehend gleich und entsprechen den formalen Vorgaben der KMK-Beschlüsse von 2003. Höchst unterschiedlich sind jedoch die Art der Implementation, Format und Umfang der Leistungsrückmeldungen, die zur Verfügung gestellten Unterstützungssysteme und das projektinterne Evaluationskonzept.

Die Ziele der Diagnose- und Vergleichsarbeiten in Baden-Württemberg sind, den Lernstand der Schüler objektiv zu dokumentieren, diagnostische Informationen zu liefern und Teil eines schulinternen und -externen Qualitätssicherungssystems zu sein (s. URL: http://www.dva-bw.de). Das dem Kultusministerium angegliederte Landesinstitut für Schulentwicklung in Stuttgart entwickelte Vergleichsarbeiten und bereitete die schulische Durchführung vor. Die Testaufgaben wurden von Teams aus erfahrenen Lehr-

kräften und Psychometrikern konzipiert und in gesonderten Vorstudien erprobt. Auf fachspezifische Kompetenzmodelle und eine Raschskalierung wurde verzichtet. Gründe hierfür konnten den offiziellen Dokumenten nicht entnommen werden.

Das Landesinstitut entwickelte schulformspezifische Tests für die Jahrgangsstufen 6 und 8. Hauptschulen mussten aus den Fächern Deutsch, Mathematik und Englisch je zwei Vergleichsarbeiten auswählen. Für Realschulen und Gymnasien waren die Kernfächer Deutsch und Mathematik verpflichtend. Zudem standen Vergleichsarbeiten für Nicht-Kernfächer zur Auswahl. Die Durchführung, Auswertung und Interpretation der Tests wurde ganz den Schulen überlassen. Das Landesinstitut stellte lediglich die Aufgabenblätter, eine Excel-Eingabemaske, die Auswertungstabellen mit landesweiten, nicht korrigierten Vergleichswerten sowie eine Handreichung für die Testdurchführung und -interpretation zur Verfügung. Die Testmaterialien konnten eine Woche vor den Durchführungsterminen Ende Juni/Anfang Juli 2007 von den Schulleitern aus dem Intranet heruntergeladen werden.

Bei der Durchführung, Auswertung und Interpretation der Daten vertraute man ganz auf die Kompetenz der Lehrkräfte. Als landesweiter Vergleichswert wurden die Ergebnisse der bereits ein Jahr zuvor durchgeführten Pilotstudie angegeben. Sozioökonomische Rahmenbedingungen der Schule und der Schüler, die Leistungsergebnisse zum großen Teil erklären können, wurden nicht in Rechnung gestellt. Den Schulen wurde im Rahmen der Selbstevaluation der Umgang mit den Rückmeldungen freigestellt. Für die Sekundarstufentests galt jedoch, dass diese benotet und als Klassenarbeit in die Jahresendnote mit einfließen mussten. Zudem wurde eine Einbindung der Vergleichsarbeitsdaten in die damals noch nicht implementierte externe Schulevaluation angekündigt. Zumindest den Schulleitern war bekannt, dass die Leistungsdaten samt den daraus abgeleiteten Schlussfolgerungen in das zukünftig zu erstellende Schulportfolio übernommen werden müssen.

In Thüringen werden jährliche Kompetenztests (siehe http://www.kompetenztest.de) von einer Arbeitsgruppe an der Universität Jena im Auftrag des Thüringer Kultusministeriums durchgeführt (Nachtigall 2005; Nachtigall/Kröhne 2006; Nachtigall 2007). Ziele sind die Analyse der Stärken und Schwächen des eigenen Unterrichts, die Rückmeldung individueller Lernstände und die Bereitstellung wissenschaftlich gesicherter Daten für die Unterrichtsentwicklung. Auf Landesebene sollen zudem Daten für die Bildungsberichterstattung bereitgestellt werden. Die Kompetenztests werden von einer multidisziplinär zusammengesetzten Arbeitsgruppe entwickelt und orientieren sich an den Anforderungsbereichen, die in den Thüringer Lehrplänen festgeschrieben sind. Ein Teil der Tests wurde in Zusammenarbeit mit anderen Bundesländern entwickelt. Jeder Teiltest basiert auf einem fachspezifischen Kompetenzmodell, das eine kriteriale Einordnung der Testwerte ermöglicht.

Die Kompetenztests wurden im Frühjahr 2007 in der Jahrgangsstufe 6 in den Fächern Deutsch, Mathematik und Englisch und in der Jahrgangsstufe 8 in Mathematik durchgeführt. Ähnlich wie in Baden-Württemberg erhielten die Schulen die Testhefte per download und waren selbst für die Durchführung und Auswertung verantwortlich. Die Dateneingabe erfolgte allerdings in einem passwortgeschützten Bereich auf der Homepage der Projektgruppe an der Universität Jena. Im Landesbericht 2007 wird zudem über ein mehrstufiges Verfahren zur Sicherung der Auswertungsobjektivität berichtet. Die

Testhefte der Schüler wurden zunächst von den Fachlehrkräften ausgewertet und gingen dann zur Zweitkorrektur an den jeweiligen Fachberater. Darüber hinaus wurden bei den Kompetenztests 2007 zufällig ausgewählte Schulen aufgefordert, die Testhefte an die Fachberater auf Schulamtsebene zu schicken. Diese Zweitauswertungen wurden zentral analysiert und Übereinstimmungswerte wurden im Landesbericht veröffentlicht.

Einer der größten Unterschiede zum baden-württembergischen Pendant ist der zeitliche und inhaltliche Umfang der Datenrückmeldung. Den Schulen wurden in drei Phasen verschiedene Leistungsindikatoren zur Verfügung gestellt. Ein erster Sofortbericht als Rückmeldung für Schüler und Eltern konnte zwei Wochen nach Eingabe der Daten heruntergeladen werden. Er enthielt die Lösungshäufigkeiten der einzelnen Aufgaben sowie Klassenergebnisse zu den einzelnen Kompetenz- und Anforderungsbereichen. Kurz nach den Sommerferien im August 2007 erhielten die Schulen die Ergebnisberichte der Klassen, die vertiefenden Ergänzungsberichte sowie die Schulberichte mit den landesweiten, SES-korrigierten Vergleichsdaten. Diese Dokumente lieferten die Datengrundlage für eine mögliche Ableitung von Unterrichts- und Schulentwicklungsmaßnahmen. Die beteiligten Lehrkräfte hatten überdies die Möglichkeit, kurz vor der Testdurchführung die Itemschwierigkeiten einzuschätzen. Der Vergleich mit den tatsächlichen Lösungshäufigkeiten in der Klasse ermöglichte eine Selbsteinschätzung der diagnostischen Kompetenz.

Das gesamte Thüringer Testkonzept wurde nach jedem Durchgang einer standardmäßigen Evaluation durch alle beteiligten Gruppen unterzogen. Das Feedback im Rahmen von Fortbildungsveranstaltungen sowie die Auswertung der Telefonhotline führten zu einer Weiterentwicklung der Eingabetechnik, der Rückmeldeformate und der Testaufgaben. Um die Ergebnisberichte effektiv nutzen zu können, werden für Lehrkräfte regelmäßig Fortbildungen und Regionalkonferenzen angeboten. Zunehmend steht hier die Nutzung und Interpretation der Daten im Vordergrund. Den Lehrkräften werden außerdem Materialien zur Weiterarbeit im Unterricht angeboten (z. B. Fehleranalysen, Lernaufgaben etc.). Die Gefahr einer gezielten Testvorbereitung wird dabei von der Projektgruppe an der Universität Jena als durchaus realistisch erkannt und sehr kritisch diskutiert.

Durch die Test- und Rückmeldesysteme in Baden-Württemberg und Thüringen werden zentrale Kriterien „guter" Vergleichsarbeiten und Leistungsrückmeldungen deutlich variiert: Berücksichtigung sozioökonomischer Hintergrundvariablen bei der Berechnung von Vergleichswerten, kriteriale Bezugsnorm durch Kompetenzmodellierung, adressatengerechte Rückmeldeformate, internes Qualitätssicherungskonzept und Fortbildungsangebote für Schulen. Die Thüringer Kompetenztests erfüllen alle diese Kriterien. Inhaltlich und bezüglich der Aufgabenformate unterscheiden sich die Vergleichsarbeiten und Kompetenztests nicht bedeutsam. Die Testaufgaben orientieren sich an den jeweiligen Lehrplänen bzw. Standards und den darin festgelegten Anforderungsbereichen. Dabei werden Reproduktions-, Anwendungs- und Transferaufgaben miteinander kombiniert. In der Regel geht es um die Prüfung grundlegender Fähigkeiten der Schüler in den zentralen Bereichen Sprache und Mathematik.

Es wird vermutet, dass sich diese Unterschiede auch auf die Akzeptanz der Tests durch Lehrer sowie die Nutzung der Leistungsdaten in den Schulen auswirken. Vor allem aus den lernpsychologischen Feedbackmodellen (Kluger/De Nisi 1996; Hattie/

Timperley 2007) lässt sich folgern, dass kriteriale und auf die Gegebenheiten der Schule angepasste Leistungsrückmeldungen zu einer intensiveren Reflexion des eigenen Unterrichts führen. Lehrkräfte in Thüringen erhalten im Vergleich zu ihren baden-württembergischen Kollegen durch zentrale Testrückmeldungen mehr relevante Hinweise über die Effektivität ihres eigenen Unterrichts. Die Rückmeldung ist somit eher in der Lage, die Aufmerksamkeit der Lehrer auf die eigentliche Aufgabe des Unterrichtens zu fokussieren. In Baden-Württemberg besteht viel eher die Gefahr, dass selbstwertdienliche Überlegungen die Diskussion über Testergebnisse dominieren.

Folgende Forschungsfragen sollen mit der vergleichenden Studie beantwortet werden:

1. Wie werden die zentralen Test- und Rückmeldesysteme in den beiden Bundesländern Baden-Württemberg und Thüringen von Lehrkräften hinsichtlich der Akzeptanz und der Nutzung für die Unterrichtsentwicklung eingeschätzt?
2. Gibt es schulische und lehrerspezifische Kontextvariablen, die diese Bewertungen beeinflussen?
3. In welchen schulinternen Gremien werden die Leistungsrückmeldungen der Vergleichsarbeiten mit welcher Intensität besprochen? Gibt es auch hier länderspezifische Differenzen?

3. Methode

Die hier berichtete Studie ist Teil einer längsschnittlich angelegten Untersuchung der Rezeption und Nutzung von Vergleichsarbeitsrückmeldungen in Baden-Württemberg (Maier 2008). Quantitative Lehrerbefragungen werden dabei mit qualitativen Interviewstudien an einzelnen Schulen gekoppelt. Im Jahr 2007 wurde die quantitative Befragung auch auf das Bundesland Thüringen ausgedehnt, um die oben beschriebenen Forschungsfragestellungen beantworten zu können.

3.1 Stichproben

In Baden-Württemberg wurden direkt nach den verpflichtenden Vergleichsarbeiten im Juni/Juli 2007 ca. 48% der öffentlichen Sekundarschulen zufällig ausgewählt und angeschrieben. Hierzu wurden die 42 Stadt- und Landkreise in Baden-Württemberg nach gymnasialen Übergangsquoten angeordnet und in jedem zweiten Kreis wurden alle öffentlichen Hauptschulen, Realschulen und Gymnasien angeschrieben. Durch dieses Auswahlverfahren wurden das Stadt-Land-Gefälle und die sozioökonomischen Differenzen zwischen den Schulamtsbezirken berücksichtigt.

Aufgrund des mehrstufigen Rückmeldeverfahrens in Thüringen konnte mit der schriftlichen Befragung erst im Oktober 2007 begonnen werden. Der Fragebogen wurde ebenfalls an die Schulleitungen einer Zufallsauswahl von 50% aller Gesamtschulen, Regelschulen und Gymnasien geschickt. Grundlage dieser Stichprobenziehung war im Gegensatz zu Baden-Württemberg eine Schulliste, aus der jede zweite Schule gezogen

wurde. Um den Rücklauf in dem kleineren Bundesland Thüringen zu erhöhen, wurden im Fragebogen zusätzlich Lehrkräfte angesprochen, die in Klasse 8 den Kompetenztest in Mathematik durchgeführt haben.

Tabelle 1 zeigt die Stichprobenstatistik aufgeschlüsselt nach Land und Schulform sowie die Rückläufe. Die Rücklaufquoten können aufgrund der Anonymität der Befragung und des zweistufigen Verteilungsverfahrens über die Schulleitungen nicht berechnet werden. Ebenso ist es nicht möglich, die Anzahl der mit zentralen Tests befassten Lehrkräfte genau zu bestimmen. Vor allem in den Hauptschulen und Regelschulen wird aufgrund des Klassenlehrersystems eine Lehrkraft in mehreren Fächern für die Durchführung der Vergleichsarbeiten bzw. Kompetenztests verantwortlich sein. Hinzu kommt, dass die Schulstatistik in Baden-Württemberg keine Aussagen über die Anzahl der Parallelklassen in Jahrgangsstufe 6 macht. Es werden lediglich die Klassen pro Schule insgesamt aufgeführt.

Auffallend ist die höhere Quote der Lehrkräfte in Thüringen, die der Schulleitung angehören. Dieser Effekt geht vor allem auf die im Vergleich zu Gymnasien klei-

Tabelle 1: Stichprobe und Rücklauf

		Angeschriebene Schulen (Anteil an Grundgesamtheit pro Schulform[1])	Klassen[4]	Rücklauf Fragebögen	Anteil Lehrerinnen	Altersschnitt (Jahre)	Mitglied Schulleitung[5]
Baden-Württemberg	Hauptschule	577 (45,9%)	860	260	62,9%	44,6	9,4%
	Realschule	212 (49,3%)	680	272	64,6%	45,8	8,1%
	Gymnasium	200 (53,1%)	740	293	50,7%	57,2	10,8%
	Gesamt	989 (47,9%)	2280	825	59,2%	45,9	9,5%
Thüringen[2]	Regelschule	120 (51,9%)	360	207	88,9%	48,4	14,0%
	Gymnasium[3]	56 (50,2%)	280	104	70,9%	46,3	8,2%
	Gesamt	167 (50,7%)	640	311	82,3%	47,7	12,0%
Gesamt		1516	2920	1136	65,7%	46,4	10,1%

Anmerkungen:

1 Datengrundlage für die Stichprobenziehung und die Stichprobenstatistik sind die aktuellen Schulstatistiken (Statistisches Landesamt Baden-Württemberg und Statistikstelle des Thüringer Kultusministeriums).

2 Die mit dem Schuljahr 2006/07 aufgehobenen Schulen in Thüringen wurden nicht mitgezählt.

3 Den Gymnasien in Thüringen wurden noch 8 in der Statistik als öffentliche Gesamtschulen ausgewiesene Schulen zugeordnet.

4 Anzahl der Klassen 6 in Baden-Württemberg und Anzahl Klassen 6 und 8 in Thüringen wurden aufgrund der durchschnittlichen Anzahl 6. Klassen pro Schule in der Stichprobe 2007 geschätzt: Regelschulen mit 1,6 Klassen pro Jahrgang; Gymnasien Thüringen mit 2,5 Klassen pro Jahrgang; Hauptschulen mit 1,5 Klassen, Realschulen mit 3,2 Klassen und Gymnasien in BW mit 3,7 Klassen pro Jahrgang.

5 Anteil der Lehrkräfte, die angeben ein Mitglied der Schulleitung zu sein.

neren Regelschulen zurück, in denen Mitglieder der Schulleitung immer noch eine umfangreiche Lehrverpflichtung haben. Für Baden-Württemberg kann man eine leichte Überrepräsentierung der Gymnasiallehrkräfte erkennen. In Thüringen sind es dagegen eher die Lehrkräfte an Regelschulen, die eifriger geantwortet haben. Als Indikator für die Schulgröße wurde nach der Anzahl der 6. Klassen gefragt. Ebenso wurde nach der Größe der Klasse gefragt, in der die Vergleichsarbeit durchgeführt wurde (Tabelle 2). Aus landesstatistischen Daten ließ sich für die angeschriebenen Schulen in Baden-Württemberg die durchschnittliche Anzahl sechster Klassen pro Schule sowie die Klassengröße schätzen. Die Schulgrößen der antwortenden Haupt- und Realschullehrkräfte entsprechen in etwa den geschätzten Größen der angeschriebenen Schulen. Eine Abweichung zeigt sich wiederum bei der Gymnasialstichprobe. Lehrkräfte aus größeren Gymnasien scheinen häufiger geantwortet zu haben.

Während die Klassengrößen der befragten Realschullehrkräfte den statistischen Schätzungen für die angeschriebenen Schulen entsprechen, haben eher Hauptschul- und Gymnasiallehrer mit kleineren Klassen geantwortet. Ein möglicher Grund dafür wäre, dass die Belastung bei größeren Klassen eher zur Ablehnung der Teilnahme an einer Befragung führen kann.

Als bedeutsame berufsbezogenene Persönlichkeitsvariable wurde die Lehrerselbstwirksamkeitserwartung erhoben. Hintergrund ist die Überlegung, dass Selbstwirksamkeit eine zentrale, berufsbezogene Disposition darstellt und eine geringe Ausprägung zu einer stärkeren Verweigerung zusätzlicher Aufgaben führt. Um zu prüfen, ob die Stichprobe diesbezüglich verzerrt ist, wurde die Skala zur Lehrerselbstwirksamkeitserwartung aus dem Modellversuch „Selbstwirksame Schule“ eingesetzt (Schwarzer/Jerusalem 1999; Schmitz/Schwarzer 2000). Für den hier vorliegenden Datensatz ergab sich eine interne Skalenkonsistenz von .78 (alpha) und mittlere Summenwerte zwischen 28,9 (BW 2007) und 29,7 (THÜR 2007). Die in Thüringen befragten Lehrkräfte geben eine signifikant höhere Selbstwirksamkeit an als ihre Kollegen in Baden-Württemberg (N=1106; F=5.919; p=0.015). Der Unterschied zum oberen Wert bei Schwarzer und Jerusalem (1999) ist ebenfalls signifikant (T=2.173; df=298; p=0.031). Aufgrund dieser Differenz zugunsten der Thüringer Lehrkräfte in der Stichprobe, wird im Ergebnisteil zu

Tabelle 2: Schul- und Klassengrößen

		Durchschnittliche Schulgröße in der Stichprobe (Anzahl 6. Klassen)	Schätzung der Schulgröße mithilfe der Schulstatistik (Anzahl 6. Klassen)	Durchschnittliche Klassengröße in der Stichprobe (Schüler)	Schätzung der Klassengröße mithilfe der Schulstatistik
BW 2007	Hauptschule	1,46	1,48	18,5	20,3
	Realschule	3,15	3,23	27,7	27,5
	Gymnasium	3,66	3,04	28,8	33,5
THÜR 2007	Regelschule	1,59	k.A.	17,8	k.A.
	Gymnasium	2,53	k.A.	22,7	k.A.

prüfen sein, inwiefern die Lehrerselbstwirksamkeit mit den Einstellungen zu zentralen Tests zusammenhängt.

Zu Beginn der Einschätzungsskalen wurden die Lehrkräfte ebenfalls gefragt, auf welchen Test in welchem Fach sie ihre Angaben beziehen. In Baden-Württemberg konnten sich die Schulen optional für Vergleichsarbeiten in weiteren Nebenfächern entscheiden. Die Angaben der befragten Lehrkräfte bezogen sich jedoch überwiegend auf die Tests in den Hauptfächern Deutsch (36,3%) und Mathematik (40,6%). Aufgrund der zu kleinen Fallzahlen in den Nebenfächern werden bei fachspezifischen Untergruppenvergleichen lediglich die Bewertungen der Testrückmeldungen in Deutsch und Mathematik miteinander verglichen. In Thüringen wurde zusätzlich noch nach den Kompetenztests in der Klassenstufe 8 gefragt, um die Rücklaufquote zu erhöhen. Die Lehrerantworten verteilen sich nahezu gleichmäßig auf alle vier Kompetenztests (Deutsch 6: 22,8%; Mathematik 6: 27,0%; Englisch 6: 23,5% und Mathematik 8: 26,0%).

3.2 Instrumente

Die eingesetzten Skalen basieren auf der grundsätzlichen Überlegung, dass Lehrkräfte die Adressaten der Vergleichsarbeiten sind und ihre Einstellung gegenüber den Tests und den Leistungsrückmeldungen von entscheidender Bedeutung für die weitere Nutzung ist (z. B. Groß Ophoff et al. 2006; Bonsen/Büchter/Peek 2006). Als Indikatoren für die Akzeptanz wurden Items zur Einschätzung des allgemeinen Nutzens der Tests, zu negativen Begleiteffekten und zur curricularen Validität (curricular alignment) verwendet.

Die im theoretischen Teil besprochenen Modelle weisen auch darauf hin, dass die Leistungsrückmeldungen nicht nur akzeptiert, sondern auch konkret nutzbare Informationen für bestimmte Tätigkeitsfelder einer Lehrkraft bereitstellen müssen (Visscher/Coe 2003; Hattie/Timperley 2007). Den administrativen Zielsetzungen in beiden Bundesländern zufolge könnten sich Hinweise für folgende Tätigkeitsfelder von Lehrkräften ergeben: Individuelle Lernstandsdiagnose, Notengebung und Selektionsdiagnose, Unterrichtsentwicklung (Curriculum und Aufgaben).

Für die Erfassung der Einstellung gegenüber zentralen Tests und der subjektiven Einschätzung der Nutzungsmöglichkeiten von Leistungsrückmeldungen wurden insgesamt 35 Items formuliert. Eine explorative Hauptkomponentenanalyse bestätigte bis auf wenige Items die theoretischen Vorüberlegungen. Mit insgesamt sieben Faktoren konnte ca. 2/3 der Gesamtvarianz erklärt werden. Mit drei Skalen wurden Einstellungen der Lehrkräfte zu den Vergleichsarbeiten erfasst:

- Allgemeine Akzeptanz zentraler Tests (6 Items, alpha = .87): Mit dieser Skala wurden die Lehrkräfte nach ihrer Meinung zur allgemeinen Bedeutung zentraler Tests für die Schule und die Weiterentwicklung des Unterrichts gefragt (z. B.: „Die Vergleichsarbeit ist für die Arbeit der Schule sehr wichtig.“). Diese Skala entspricht der von Ditton und Merz (2000) entwickelten Skala „Einstellungen gegenüber zentralen Tests“.

- Test als Belastung (4 Items, alpha = .79): Diese Skala setzt sich aus vier Aussagen zu möglichen negativen Auswirkungen zentraler Tests innerhalb der Schule zusammen (z. B.: „Die Vergleichsarbeit übt zusätzlichen Druck auf Schulen und Lehrer aus.“).
- Curriculare Validität der Vergleichsarbeit (4 Items, alpha = .84): Die Lehrkräfte wurden hier gefragt, inwiefern die zentralen Leistungsmessungen mit den Vorgaben in den Lehrplänen bzw. Bildungsstandards übereinstimmen (z. B.: „Die Vergleichsarbeit stimmt in ihren Teilbereichen mit der Gewichtung der Lerninhalte und Kompetenzen im Bildungsplan überein.“).

Mit weiteren vier Skalen wurde erhoben, inwiefern die Vergleichsarbeitsrückmeldungen als zusätzliche, nützliche Hinweise innerhalb bestimmter Bereiche des professionellen Handelns angesehen werden:

- Diagnostische Hinweise (5 Items, alpha = .89): Mit dieser Skala wurde erfasst, in welchem Maße die Rückmeldungen als zusätzliche diagnostische Informationen wahrgenommen werden und als solche auch nutzbar sind (z. B.: „Die Vergleichsarbeiten geben mir zusätzliche diagnostische Hinweise.“). Auch wenn die Verwendung von Vergleichsarbeitsdaten für individualdiagnostische Zwecke zum Teil abgelehnt, aber zumindest stark eingeschränkt wird (Nachtigall 2007), gehört die Rückmeldung individueller Lernstände zu den zentralen Zielen der jeweiligen Testsysteme.
- Hinweise für die Notengebung (5 Items, alpha =.80): Hier wurden Lehrkräfte gefragt, ob die Vergleichsarbeitsrückmeldungen die eigene Leistungsmessung und Notengebung bestätigen oder in Frage stellen können (z. B.: „Die Vergleichsarbeiten regen zum Nachdenken über den eigenen Bewertungsmaßstab an.“).
- Hinweise für zukünftige Übungen/Wiederholungen (3 Items, alpha = .90): Mit dieser Skala wurde gefragt, ob die Auswertung der Vergleichsarbeiten bestimmte Hinweise für zukünftige Wiederholungen oder Übungen gibt (z. B.: „Die Vergleichsarbeit gibt mir zusätzliche Hinweise, welche Aufgabenstellungen in Zukunft besser geübt werden müssen.“).
- Hinweise auf inhaltliche Änderungen (4 Items, alpha = .77): Eine letzte Skala erfasst die Einschätzung der Lehrkräfte, inwiefern die Vergleichsarbeit Impulse für inhaltliche Veränderungen des Unterrichts geben kann (z. B.: „Die Vergleichsarbeit gibt mir zusätzliche Hinweise, ob die Reihenfolge der behandelten Stoffgebiete geändert werden sollte.“).

Um das Ausmaß und die subjektiv wahrgenommene Qualität der schulinternen Diskussion über Vergleichsarbeitsrückmeldungen abschätzen zu können, wurde ebenfalls gefragt, an welchen Orten wie systematisch über die Testergebnisse geredet wird. Handelte es sich um einen informellen Austausch oder eine systematische Diskussion?

Bei der Erfassung der Intensität und Qualität der schulinternen Diskussion ergibt sich ein zusätzliches Messproblem. Die Fragebögen wurden jeweils kurz nach der letzten, für die Lehrer relevanten Rückmeldung an die Schulen gesandt. In Baden-Württemberg war dies direkt nach der Testdurchführung, weil die Daten sofort in die mitgelieferte Excel-Tabelle eingegeben wurden und den Lehrkräften alle Rückmeldeinformationen sofort zur Verfügung standen. In Thüringen war dies gegen Ende Oktober, als der ausführliche Ergebnisbericht vorlag. Der Fragebogenrücklauf erstreckte

sich in beiden Bundesländern über mehrere Wochen. Hinzu kommt, dass vollkommen unklar ist, in welchem zeitlichen Abstand Besprechungen und Konferenzen zu den Testergebnissen stattfinden werden. Dies bedeutet, dass Lehrkräfte, die den Fragebogen recht zügig ausgefüllt und zurückgesendet haben, die schulinterne Diskussion eventuell anders bewerten.

Dieser Fehler muss bei der Ergebnisinterpretation berücksichtigt werden. Es wird allerdings angenommen, dass die Verzerrungen nicht übermäßig sein werden. Zum einen tritt der Fehler in beiden Bundesländern vermutlich in der gleichen Ausprägung auf. Der Fragebogen lag den Schulen immer dann vor, wenn die eigentliche Diskussion beginnen konnte. Zum anderen werden die Lehrkräfte sehr wahrscheinlich die Situation an ihrer Schule recht gut einschätzen können und die möglichen Diskussionen über Testergebnisse bei der Beantwortung der Fragen antizipieren. Dies trifft vor allem auf die Daten der Erhebung 2007 zu. Hier kann in beiden Bundesländern bereits auf Erfahrungswerte aus den vergangenen Schuljahren zurückgegriffen werden.

4. Ergebnisse

4.1 Akzeptanz und Nutzung im Ländervergleich

Zunächst werden die Skalenwerte zur Akzeptanz und zu Nutzungsoptionen für die Erhebung 2007 im Ländervergleich Baden-Württemberg vs. Thüringen dargestellt. Da die Schulformen in beiden Bundesländern nicht vergleichbar sind, wird nicht schulformspezifisch differenziert.

Tabelle 3 zeigt die deskriptiven Statistiken der gerechneten T-Tests. Alle Einschätzungsvariablen unterscheiden sich in Abhängigkeit des Bundeslandes hochsignifikant voneinander. Auch die Varianzen der Variablen in den beiden Ländern unterscheiden sich deutlich. Dies reflektiert sehr wahrscheinlich schulformspezifische Differenzen in Baden-Württemberg, die bereits in der Befragung im Jahr 2006 sehr deutlich zum Vorschein kamen (Maier 2008). Von Hauptschullehrkräften werden die Vergleichsarbeiten wesentlich positiver wahrgenommen als von Lehrkräften an Realschulen und Gymnasien. Einen Schulformunterschied in Thüringen gibt es lediglich bei einer Variablen: Lehrer an Regelschulen sehen in den Rückmeldungen wesentlich mehr Hinweise für zukünftige Wiederholungen und Übungen ($F=8{,}62$; $p<0.01$).

Welche Mittelwertsunterschiede ergaben sich? Bis auf eine Ausnahme werden von den Thüringer Lehrkräften die zentralen Tests deutlich besser bewertet als von den in Baden-Württemberg befragten Lehrkräften. Lediglich bei Hinweisen für die Notengebung schneiden die baden-württembergischen Vergleichsarbeiten besser ab. Dies kann zunächst einmal mit der Tatsache zusammenhängen, dass die Testwerte der Vergleichsarbeiten direkt in eine Note umgerechnet werden können und diese per Erlass in die Jahresendnote mit einfließen muss. In Thüringen steht die Notengebung im Hintergrund und wird den Schulen überlassen. Die Lehrkräfte sollen sich zunächst einmal mit dem fairen Schulvergleich (korrigierte Punktwerte) und der Aufgabenanalyse beschäftigen. Dieses Ergebnis zeigt, dass der eingesetzte Fragebogen auf Differenzen der Testsysteme sensibel reagiert.

Tabelle 3: Ländervergleich: Einschätzung der Akzeptanz und Nutzung zentraler Tests (T-Tests)

	Bundesland	N	Mittelwert [1]	SD	Signifikanz
Allg. Akzeptanz von Vergleichsarbeiten	BW	816	2,68	1,08	$p < 0.001$
	THÜR	309	3,47	0,90	
Vergleichsarbeiten als Belastung	BW	816	2,69	1,01	$p < 0.001$
	THÜR	309	2,06	0,87	
Lehrplanvalidität der Vergleichsarbeit	BW	816	2,85	0,94	$p < 0.001$
	THÜR	309	3,45	0,84	
Diagnostische Hinweise	BW	826	2,59	1,06	$p < 0.001$
	THÜR	312	3,19	0,94	
Hinweise Notengebung	BW	827	2,81	1,00	$p < 0.01$
	THÜR	311	2,62	0,85	
Hinweise auf zukünftige Wiederholungen und Übungen	BW	819	2,93	1,22	$p < 0.001$
	THÜR	310	3,60	0,97	
Hinweise auf inhaltliche Änderungen	BW	821	2,08	0,89	$p < 0.001$
	THÜR	310	2,32	0,90	

Anmerkungen:
1 Die Einzelitems der jeweiligen Skalen variieren zwischen dem Minimalwert 1 (Ablehnung) und dem Maximalwert 5 (Zustimmung).

Die Unterschiede in den einzelnen Bewertungsdimensionen sind zum Teil erheblich. In Thüringen liegen die allgemeine Akzeptanz und die eingeschätzte curriculare Validität zentraler Tests deutlich über dem semantischen Median von 3. Die Kompetenztests werden auch als weitaus weniger belastend für die Arbeit an Schulen wahrgenommen als die Vergleichsarbeiten in Baden-Württemberg. Ebenso wird die Nutzung der Testergebnisse für Lerndiagnosen und die Unterrichtsentwicklung höher eingeschätzt.

Lediglich der Hinweischarakter für zukünftige inhaltliche Änderungen wird in beiden Bundesländern gleich niedrig bewertet (Boden-Effekt). Sowohl die Kompetenztests als auch die Vergleichsarbeiten in Baden-Württemberg beziehen sich auf die jeweiligen Lehrpläne und prüfen grundlegende Fähigkeiten ab. Die Wahrscheinlichkeit, dass die Aufgabenstellungen durch den Stoffverteilungsplan der Lehrkräfte bzw. das verwendete Schulbuch abgedeckt werden, ist somit sehr hoch. Dies könnte erklären, dass Hinweise auf inhaltliche Veränderungen eher nicht wahrgenommen werden.

Die Stichprobenanalyse ergab Unterschiede zwischen Baden-Württemberg und Thüringen, die sich möglicherweise auf die Einschätzung der zentralen Tests auswirken können. In Thüringen sind die Schulen und Klassen im Durchschnitt kleiner, weniger Schüler haben einen Migrationshintergrund, mehr Mitglieder der Schulleitung haben auf die Befragung geantwortet und die Lehrkräfte berichten günstigere Selbstwirksamkeitserwartungen. Aus diesem Grund wurde mit Regressionsanalysen geprüft, in welchem Maße diese Drittvariablen die Einschätzung der Vergleichsarbeitsrückmeldungen beeinflussen. Ebenso wird die Variable „getestetes Unterrichtsfach“ (Deutsch oder Mathematik) mit eingebaut, um den Effekt unterschiedlicher Fachkulturen auf die Einschätzung

Tabelle 4: Effekte weiterer Kontextvariablen auf die Lehrereinschätzung der Vergleichsarbeiten (Standardisierte Regressionskoeffizienten)

	Allg. Akzeptanz	Belastung	Curriculare Validität	Diagn. Hinweise	Hinweise Notengebung	Hinweise Wiederholungen	Hinweise Inhalte
Bundesland: BW- THÜR [1]	.26 ***	-.13 **	.21 ***	.21 ***	-.14 **	.30 ***	.10 *
Schulform: Gymnasium[2]	n.s.	n.s.	-.11 **	n.s.	.10 *	-.12 **	-.13 **
Schulgröße (Anzahl 6. Klassen)	n.s.	n.s.	-.11 *	n.s.	-.20 ***	n.s.	n.s.
Klassengröße (Anzahl Schüler)	-.11 *	.21 ***	n.s.	n.s.	n.s.	.12 *	.11 *
Anzahl Schüler mit Migrationshintergrund	n.s.	n.s.	n.s.	n.s.	.09 *	n.s.	n.s.
Lehrerselbstwirksamkeitserwartung	.08 *	n.s.	.08 *	.10 **	n.s.	n.s.	n.s.
Lehrer ist Mitglied der Schulleitung: ja	n.s.	-.08 *	n.s.	n.s.	n.s.	n.s.	n.s.
Schulfach: Deutsch – Mathematik [3]	.15 ***	n.s.	.11 **	.11 **	.12 **	.14 ***	.10 *
N	673	673	672	675	676	673	675
R^2	16,3%	13,0%	14,5%	10,7%	7,3%	13,7%	4,4%

Anmerkungen: *** p <.001, ** p <.01, * p <.05

1 Fragebögen aus Baden-Württemberg wurden mit 1, Fragebögen aus Thüringen mit 2 kodiert.

2 Aufgrund unterschiedlicher Schulsysteme wurden für die Regressionsanalysen die Fragebögen aus Regelschulen in Thüringen und Haupt- und Realschulen in Baden-Württemberg mit 1 kodiert. Die Fragebögen von Gymnasiallehrern wurden in beiden Bundesländern mit 2 kodiert.

3 Fragebögen zu einem zentralen Test im Fach Deutsch wurden mit 1 kodiert, Fragebögen zu einem zentralen Test in Mathematik wurden mit 2 kodiert.

der Vergleichsarbeiten kontrollieren zu können. Um stabilere Ergebnisse zu erhalten, werden die Angaben zu den Mathematik-Kompetenztests in Klasse 6 und Klasse 8 zusammengefasst. Dies ist möglich, da mit t-Tests geprüft wurde, dass es keine signifikanten Unterschiede in der Einschätzung der Thüringer Kompetenztests Mathematik nach Jahrgangsstufe gibt.

Aufgrund der unterschiedlichen Schulgliederung können die Schulformeffekte in den Regressionsanalysen nicht direkt berechnet werden. Deshalb wurden jeweils Haupt- und Realschulen in Baden-Württemberg sowie Regelschulen in Thüringen gegen Gymnasien verglichen. Vor allem in Baden-Württemberg zeigte eine frühere Befragung, dass Lehrkräfte an Gymnasien den zentralen Tests gegenüber grundsätzlich kritischer eingestellt sind (Maier 2008).

Zunächst einmal bestätigen die Regressionsanalysen den Länderunterschied in der Lehrereinschätzung der zentralen Tests. Unabhängig von Schulform, Schulfach, schulischen Kontextmerkmalen, Mitgliedschaft in der Schulleitung und der Lehrerselbstwirksamkeitserwartung werden die Kompetenztests in Thüringen insgesamt eher akzeptiert und für nützlicher angesehen als die Vergleichsarbeiten in Baden-Württemberg.

Schulformeffekte treten nur bei einem Teil der Variablen auf. Gymnasiallehrkräfte schätzen die Lehrplanvalidität der Tests als geringer ein. Ebenso sehen sie weniger nützliche Hinweise auf Veränderungen des Curriculums und der zu übenden Aufgabenstellungen. Dafür kann man an Gymnasien den Vergleichsarbeiten und Kompetenztests eher einen selektionsdiagnostischen Nutzen abgewinnen.

Die Größe einer Schule macht sich lediglich bei der curricularen Validität und den Hinweisen für die Notengebung bemerkbar. In beiden Fällen werden diese Aspekte der Vergleichsarbeiten an größeren Schulen eher negativ bewertet. Als mögliche Erklärung hierfür könnte man annehmen, dass an Schulen mit mehr Parallelklassen ohnehin genug Vergleichsmöglichkeiten für die eigene Leistungsbeurteilung bestehen und Lehrkräfte somit nicht auf zentrale Leistungstests angewiesen sind.

Auch die Klassengröße spielt nur bei einem Teil der untersuchten Variablen eine Rolle. Sehr deutlich ist allerdings der Effekt der Klassengröße auf die vom Lehrer wahrgenommene Belastung durch eine zentrale Leistungsmessung: Bei großen Klassen ist die Durchführung und sehr genau vorgeschriebene Auswertung eines zusätzlichen Tests ein entsprechend großer Arbeitsaufwand und könnte somit zu einer eher ablehnenden Haltung beitragen. Ebenso könnte die mit größeren Klassen verbundene, generelle Arbeitsbelastung auf die Beurteilung der Vergleichsarbeiten ausstrahlen. Dagegen sehen Lehrer mit größeren Klassen wiederum eher einen Nutzen der Tests für die zukünftige Planung und Anordnung von Inhalten und Wiederholungseinheiten.

Der Anteil der Schüler mit Migrationshintergrund in den getesteten Schulklassen ist in Baden-Württemberg wesentlich höher als in Thüringen. In den Regressionsanalysen zeigt sich jedoch, dass diese Variable nur an einer Stelle eine geringe Auswirkung auf die Lehrereinschätzung hat. Lehrkräfte in Klassen mit vielen nicht-deutschsprachigen Schülern nutzen die Rückmeldungen eher als Hinweise für die eigene Notengebung als Lehrkräfte in homogeneren Klassen. Die Vergleichsarbeiten scheinen allerdings nicht sehr nützlich zu sein, um die Individualdiagnose in heterogenen Klassen zu unterstützen.

Die Lehrerselbstwirksamkeitserwartung hat geringe Effekte, vor allem auf die Akzeptanzvariablen. Lehrer mit einer höheren Selbstwirksamkeitserwartung fühlen sich weniger belastet und schätzen die Lehrplanvalidität sowie den allgemeinen und diagnostischen Nutzen der Vergleichsarbeiten besser ein. Noch geringere Auswirkungen auf die Testeinschätzung hat die Tatsache, ob eine Lehrkraft zugleich noch Mitglied der Schulleitung ist. Wenn Schulleiter oder ihre Stellvertreter selbst an der Testdurchführung beteiligt sind, nehmen sie lediglich eine geringere Belastung wahr. Dies könnte ein

Hinweis darauf sein, dass ein Teil des Belastungsempfindens auf die mögliche Kontrolle durch die Schulleitung zurückzuführen ist. Gehört man selbst der Schulleitung an, reduziert sich diese Belastung.

Das getestete Schulfach hat eine durchgängige Auswirkung auf die erfasste Lehrereinschätzung. Die zentralen Mathematiktests werden in allen Dimensionen besser bewertet als die zentralen Arbeiten im Fach Deutsch. Dies gilt sowohl für die allgemeine Einstellung als auch für die Nutzungsmöglichkeiten der Testrückmeldungen.

4.2 Orte und Qualität der schulinternen Diskussion über Leistungsrückmeldungen

Die Einschätzungen zur Qualität der Diskussion in den einzelnen Gremien wurden durch ein sehr grobes Rating erfasst: (1) keine Diskussion; (2) informeller Austausch; (3) systematische Diskussion. Den Lehrkräften wurde diese Auswahl für eine Reihe möglicher Gremien bzw. Personengruppen vorgelegt. Fehlende Werte bedeuten entweder, dass es diese Art von Gremium an der Schule nicht gibt oder die befragte Lehrkraft zu möglichen Diskussionen in diesem Gremium keine Aussagen machen kann. Tabelle 5 stellt zwei Informationen im Ländervergleich dar: Zunächst kann der prozentuale Anteil der Lehrkräfte verglichen werden, die überhaupt von einer Diskussion in dem jeweiligen Gremium berichten. Danach wird der mittlere Wert der Einschätzung der Qualität dieser Diskussion über Testergebnisse aufgeführt. In der letzten Spalte wird die Wahrscheinlichkeit angegeben, dass sich die Qualitätseinschätzungen nach Land zufällig unterschieden.

Sowohl in Baden-Württemberg als auch in Thüringen diskutiert ein großer Teil der Lehrkräfte auf einer eher informellen Ebene mit ausgewählten Kollegen über die Ergebnisse der zentralen Tests. Auch die Häufigkeit einer Diskussion mit Parallelkollegen ist in beiden Bundesländern gleich hoch. Die Thüringer Lehrkräfte schätzen die Systematik dieser Gespräche mit Parallelkollegen jedoch deutlich höher ein.

Der relativ niedrige Anteil an Lehrkräften, die überhaupt Diskussionen in Jahrgangsstufenkonferenzen bewerten, muss durch institutionelle Bedingungen relativiert werden. Beispielsweise finden in kleinen Haupt- und Realschulen solche Gremiensitzungen überhaupt nicht statt. Dagegen kann man davon ausgehen, dass es in sämtlichen Schulformen Klassen- und Fachkonferenzen geben sollte. Die befragten Lehrer in Thüringen bewerten vor allem die Gespräche in Fachkonferenzen überwiegend als systematisch. Dies ist ein deutlicher Unterschied zu der Bewertung der baden-württembergischen Lehrkräfte, die Diskussionen über Testrückmeldungen in Fachkonferenzen als eher informell einschätzen.

Ebenso deutlich sind die Länderunterschiede in Bezug auf Gespräche über Testergebnisse in der Gesamtlehrerkonferenz. Hier sind die prozentualen Anteile wesentlich geringer als bei Fachkonferenzdiskussionen. Während in Thüringen die Einschätzung eher in Richtung informellem Austausch geht, berichten die Lehrer in Baden-Württemberg überwiegend, dass in den Gesamtlehrerkonferenzen nicht über die Vergleichsarbeiten gesprochen wurde. In Schulkonferenzen scheint die Thematik in beiden Bundesländern nur wenig beachtet zu werden. Dagegen finden sowohl in Baden-Württemberg als auch in Thüringen in der Regel Gespräche mit der Schulleitung statt, die wiederum von

Tabelle 5: Ort und Qualität der schulinternen Diskussion im Ländervergleich

	Baden-Württemberg (N = 825)			Thüringen (N = 311)			Differenz
	N	Anteil[1]	Mittelwert	N	Anteil[1]	Mittelwert	
Ausgewählte Kollegen	651	78.9%	2,20	218	70.1%	2,18	n.s.
Parallelkollegen	691	83.8%	2,33	248	79.7%	2,47	***
Klassenkonferenz	522	63.3%	1,75	192	61.7%	2,01	***
Jahrgangsstufenkonferenz	386	46.8%	1,39	136	43.7%	1,66	***
Fachkonferenz	550	66.7%	2,11	268	86.2%	2,57	***
Gesamtlehrerkonferenz	437	53.0%	1,46	173	55.6%	1,89	***
Schulkonferenz	368	44.6%	1,15	130	41.8%	1,40	***
Schulleitung	616	74.7%	2,02	242	77.8%	2,23	***
Schüler	703	85.2%	2,35	284	91.3%	2,54	***
Eltern	580	70.3%	1,79	243	78.1%	2,09	***

Anmerkungen: *** p <.001, ** p <.01, * p <.05

1 Prozentualer Anteil der Lehrer, die zum jeweiligen Gremium bzw. zur jeweiligen Personengruppe Angaben gemacht haben.

2 Mittelwert der eingeschätzten Qualität der Diskussion im jeweiligen Gremium bzw. mit der jeweiligen Personengruppe mit folgendem Wertebereich: 1 = „keine Diskussion“, 2 = „informelle Gespräche“, 3 = „systematische Diskussion“.

3 Signifikanzniveau der Differenz zwischen beiden Mittelwerten der Qualitätseinschätzung (t-Test).

Thüringer Lehrern als systematischer eingeschätzt werden. Gleiches gilt für die Besprechung der Testergebnisse mit Schülern und Eltern.

5. Diskussion

Die Ergebnisse zeigen deutliche Länderunterschiede in der Akzeptanz und Nutzung der Vergleichsarbeitsdaten auf Lehrerebene (Forschungsfrage 1). Die Kompetenztests in Thüringen werden eher akzeptiert, weniger als Belastung angesehen und haben für die Lehrkräfte einen höheren lerndiagnostischen Hinweischarakter. In Baden-Württemberg dagegen geben Vergleichsarbeitsrückmeldungen eher Hinweise für die Notengebung.

Lediglich das getestete Schulfach beeinflusst die Testakzeptanz und die Testnutzung durchgehend (Forschungsfrage 2). Dieses Ergebnis reflektiert die in der Literatur diskutierte „Testfreundlichkeit“ des Faches Mathematik (Blum et al. 2005; Tresch 2007; Sill/Sikora 2007). In Mathematik gleichen Testaufgaben eher den im Unterricht einge-

setzten Lernaufgaben und im Vergleich zum Fach Deutsch besteht ein größerer Konsens über die Möglichkeit, mit einfach zu korrigierenden Aufgaben die Basisfertigkeiten des Faches einigermaßen zuverlässig zu prüfen. Weitere Merkmale der Schule, der Klasse und des Lehrers haben nur partiell Auswirkungen auf die Lehrereinschätzungen von Vergleichsarbeiten. Die zusätzliche Varianzaufklärung mit den eingesetzten Kontextmerkmalen ist insgesamt gering. Mit den Regressionsanalysen konnte ebenfalls geprüft werden, dass diese deutlichen Länderunterschiede bestehen bleiben, wenn man die erfassten Kontextfaktoren in Rechnung stellt.

Die Länderunterschiede wiederholen sich auch auf Schulebene (Forschungsfrage 3). Die Besprechung der zentralen Leistungsdaten in schulinternen Gremien wird von Lehrkräften in Thüringen als durchgehend systematischer eingeschätzt. Unabhängig vom Ländervergleich zeigt sich auch, dass in Gesprächen mit Parallelkollegen, Fachkollegen und Schülern am ehesten systematisch über die Testrückmeldungen diskutiert wird. Ebenfalls werden in beiden Bundesländern regelmäßige Gespräche mit der Schulleitung über diese Thematik geführt. Die geringe Resonanz der Vergleichsarbeiten in Gesamtlehrerkonferenzen lässt jedoch darauf schließen, dass eine Nutzung der zentralen Leistungsdaten für Entwicklungs- oder Diskussionsprozesse in der Einzelschule – zumindest in Baden-Württemberg – noch kaum erkennbar ist.

Grundsätzlich stellt sich nun die Frage, welche Merkmale der Test- und Rückmeldesysteme in beiden Ländern für die nachgewiesenen Differenzen in der Lehrerwahrnehmung verantwortlich gemacht werden können. Im Grunde genommen wurde ein ganzes Bündel an Merkmalen guter Vergleichsarbeiten und Rückmeldungen als Quasi-Experimentalbedingung geprüft. Aus diesem Grund lassen sich an dieser Stelle lediglich sehr breite, weiterführende Hypothesen entwickeln, die mögliche Wirkungen der Testsysteme auf die Lehrerwahrnehmung und schulinterne Nutzung beschreiben.

Eine wesentliche Differenz zwischen beiden Bundesländern ist die Art der Rückmeldung von Leistungsdaten. Die Rückmeldung SES-korrigierter Landesmittelwerte, die Aufgliederung der Klassenwerte in fachspezifische Kompetenzen und die Verknüpfung der Rückmeldungen mit weiterführenden Fehleranalysen unterstützen die Fokussierung der Aufmerksamkeit auf den vorausgehenden Unterricht. In einem lernpsychologischen Sinne wird damit das Unterrichten als eigentliche Aufgabe des Lehrers mit der Rückmeldung verknüpft (z. B. Hattie/Timperley 2007). Mögliche Veränderungshinweise in der Rückmeldung können somit auch veränderungswirksam werden. In Baden-Württemberg wurde den Lehrkräften dagegen lediglich ein sozialer Vergleichsmaßstab mit einer landesweiten Pilotierungsstichprobe angeboten. Auch auf die Bereitstellung fairer Vergleichsdaten wurde zugunsten einer Sofortauswertung verzichtet. Dies scheint die Akzeptanz- und Nutzbarkeitseinschätzungen eher beeinträchtigt zu haben.

Von baden-württembergischen Lehrkräften wird die Nutzung der Testrückmeldungen für die Notengebung höher eingeschätzt als von Lehrkräften in Thüringen. Auch dieses Ergebnis lässt sich auf einen Unterschied im Rückmeldeformat zurückführen. Die baden-württembergischen Vergleichsarbeiten werden kurz vor Ende des 6. Schuljahres geschrieben und fließen als benotete Klassenarbeit in die Jahresendnote ein. In Thüringen wird dagegen kein offizieller Notenumrechnungsschlüssel angeboten und die

Bewertung des Kompetenztests als zusätzliche schriftliche Note ist optional und muss von den Schulen entschieden werden. Diese Regelung fokussiert die Aufmerksamkeit der baden-württembergischen Lehrkräfte viel stärker auf selektionsdiagnostische Aspekte der Testrückmeldung als dies in Thüringen der Fall ist. In den Lehrereinschätzungen spiegelt sich dies entsprechend wieder.

Differenzen bezüglich der Testakzeptanz und der Qualität der Diskussion über Testergebnisse in schulischen Gremien lassen sich vermutlich sehr stark auf die unterschiedliche Einbindung in umfangreichere Reform- und Innovationsprozesse zurückführen (Modellkomponente Implementation bei Visscher/Coe 2003). In Thüringen werden die Kompetenztests zusammen mit weiteren Instrumenten der Selbstevaluation (z. B. Schüler beurteilen den Unterricht) entwickelt und den Schulen in einer konsistenten Form präsentiert. Gleichzeitig wird von der gleichen Projektgruppe ein umfangreiches Fortbildungsangebot bereitgestellt, das die Implementation der zentralen Tests begleitet und zunehmend auf die schulinterne Datennutzung abzielt. In Baden-Württemberg dagegen stehen die Vergleichsarbeiten relativ unverbunden neben anderen Elementen der aktuellen Schul- und Unterrichtsreform.

Ein weiterer Grund für die nur mangelhafte schulinterne Diskussion der Leistungsdaten in Baden-Württemberg liegt natürlich in der eher geringen Akzeptanz und Nutzung auf Lehrerebene, die der schulinternen Diskussion zeitlich vorausgehen muss (siehe Modelle bei Helmke/Hosenfeld 2005; Tresch 2007). Ebenso korrespondiert dieses Ergebnis mit weiteren Befunden der Rezeptionsforschung. Die Einschätzung der Brauchbarkeit von Vergleichsarbeiten und Leistungsrückmeldungen durch Lehrkräfte wirkt sich auf die Ableitung von Entwicklungsmaßnahmen auf Schulebene aus (z. B. Bonsen/Büchter/Peek 2006).

Wenn auch noch etliche Detailfragen zu klären sind, zeigt dieser Vergleich letztendlich wie unterschiedlich Lehrer und Schulen auf bildungspolitische Reform elemente reagieren können. Ein an wissenschaftlichen Standards orientiertes Test- und Rückmeldesystem wie die Kompetenztests in Thüringen konnte Vertrauen innerhalb der Lehrerschaft aufbauen und eine schulinterne Nutzung der Rückmeldedaten fördern. Für die weitere Erforschung der Effekte einzelner Komponenten der Vergleichsarbeiten könnten vergleichende bzw. quasi-experimentelle Studien innerhalb einzelner Bundesländer hilfreich sein. Beispielsweise könnte man bestimmte Bedingungen der Rückmeldeformate systematisch variieren und die Reaktionen in Schulen vergleichen.

Literatur

Altrichter, H./Heinrich, M. (2006): Evaluation als Steuerungsinstrument im Rahmen eines „neuen Steuerungsmodells“ im Schulwesen. In: Böttcher, W./Holtappels, H.-G./Brohm, M. (Hrsg.): Evaluation im Bildungswesen - Eine Einführung in Grundlagen und Praxisbeispiele. – Weinheim, S. 52–64.

Baeriswyl et al. 2006 = Baeriswyl, F./Wandeler, C./Trautwein, U./Oswald, K. (2006): Leistungstest, Offenheit von Bildungsgängen und obligatorische Beratung der Eltern. In: Zeitschrift für Erziehungswissenschaft, 9. Jg., H. 3, S. 371–392.

Bailey, K. (1996): Working for washback: a review of the washback concept in language testing. In: Language Testing, Vol. 13(3), pp. 257–279.

Blum et al. 2005 = Blum, W./Drüke-Noe, C./Leiß, D./Wiegand, B./Jordan, A. (2005): Zur Rolle von Bildungsstandards für die Qualitätsentwicklung im Mathematikunterricht. In: Zentralblatt für Didaktik der Mathematik (ZDM), 37(4), S. 267–274.

Bonsen, M./Büchter, A./Peek, R. (2006): Datengestützte Schul- und Unterrichtsentwicklung. Bewertungen der Lernstandserhebungen in NRW durch Lehrerinnen und Lehrer. In: Bos, W./Holtappels, H.-G./Pfeiffer, H./Rolff, H. G./Schulz-Zander, R. (Hrsg.): Jahrbuch der Schulentwicklung, Bd. 14. – Weinheim, S. 125–148.

Büchter, A./Leuders, T. (2005): From students' achievement to the development of teaching: requirements for feedback in comparative tests. In: Zentralblatt für Didaktik der Mathematik (ZDM), 37(4), S. 324–334.

Cheng, L./Curtis, A. (2004): Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning. In: Cheng, L./Watanabe, Y./Curtis, A. (Eds.): Washback in Language Testing. Research Contexts and Methods. – London, pp. 3–17.

Ditton, H./Merz, D. (2000): Qualität von Schule und Unterricht - Kurzbericht über erste Ergebnisse einer Untersuchung an bayerischen Schulen. – URL: http://www.quassu.net/Bericht1.pdf (Download: 26.03.2008).

Groß Ophoff et al. 2006 = Groß Ophoff, J./Koch, U./Hosenfeld, I./Helmke, A. (2006): Ergebnisrückmeldung und ihre Rezeption im Projekt VERA. In: Kuper, H./Schneewind, J. (Hrsg.): Rückmeldung und Rezeption von Forschungsergebnissen. – New York, S. 19–40.

Hattie, J./Timperley, H. (2007): The Power of Feedback. In: Review of Educational Research, Vol. 77/1, pp. 81–112.

Helmke, A./Hosenfeld I. (2005): Standardbezogene Unterrichtsevaluation. In: Brägger, G./Bucher, B./Landwehr, N. (Hrsg.): Schlüsselfragen zur externen Schulevaluation. – Bern, S. 127–151.

Hosenfeld, I./Schrader, F.-W./Helmke, T. (2006): Von der Rezeption zur Ergebnisrückmeldung: Leistungsevaluation im Spannungsfeld von System-Monitoring und Schulentwicklung. In: Hosenfeld, I./Schrader, F.-W. (Hrsg.): Schulische Leistung – Grundlagen, Bedingungen, Perspektiven. – Münster, S. 289–313.

Hulpia, H./Valcke, M. (2004): The Use of Performance Indicators in a School Improvement Policy: The Theoretical and Empirical Context. In: Evaluation and Research in Education, Vol. 18/1&2, pp. 102–120.

Keller, F./Moser, U. (2006): Check 5. Schlussbericht 2006 zuhanden des Departements Bildung, Kultur und Sport des Kantons Aargau. Vervielf. Ms. Zürich: KBL 2006.

Klug, C./Reh, S. (2000): Was fangen die Schulen mit den Ergebnissen an? Die Hamburger Leistungsvergleichsstudie aus der Sicht ‚beforschter' Schulen. In: Pädagogik, Heft 12, S. 16–21.

Kluger, A. N./De Nisi, A. (1996): The effects of Feedback Interventions on performance: A historical review, a meta-analysis, and a preliminary Feedback Intervention Theory. In: Psychological Bulletin, Vol. 119/2, pp. 254–284.

Kohler, B. (2004): Zur Rezeption externer Evaluation durch Lehrkräfte, Eltern sowie Beamte der Schulaufsicht. In: Empirische Pädagogik, Heft 18/1, S. 18–39.

Lorenz, J. H. (2005): Zentrale Lernstandsmessung in der Primarstufe: Vergleichsarbeiten Klasse 4 (VERA) in sieben Bundesländern. In: Zentralblatt für Didaktik der Mathematik (ZDM), Bd. 37, H. 4, S. 317–324.

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

Moser, U. (2003): Klassencockpit im Kanton Zürich – Ergebnisse einer Befragung von Lehrerinnen und Lehrern der 6. Klassen über ihre Erfahrungen im Rahmen der Erprobung von Klassencockpit im Schuljahr 2002/03. Bericht zuhanden der Bildungsdirektion des Kantons Zürich. – URL: http://www.lehrmittelverlag.ch/downloads/dateien/Evaluation%20Klassencockpit.pdf (Download: 26.03.2008).

Nachtigall, C. (Hrsg.) (2005): Landesbericht – Thüringer Kompetenztest 2005. – Friedrich-Schiller-Universität Jena.

Nachtigall, C. (Hrsg.) (2007): Landesbericht – Thüringer Kompetenztest 2007. – Friedrich-Schiller-Universität Jena.

Nachtigall, C./Kröhne, U. (2006): Methodische Anforderungen an schulische Leistungsmessung – auf dem Weg zu fairen Vergleichen. In: Kuper, H./Schneewind, J. (Hrsg.): Rückmeldung und Rezeption von Forschungsergebnissen. – Berlin, S. 59–74.

Peek, R. (2004): Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik (QuaSUM) – Klassenbezogene Ergebnisrückmeldung und ihre Rezeption in Brandenburger Schulen. In: Empirische Pädagogik, Bd. 18, H. 1, S. 82–114.

Peek, R./Dobbelstein, P. (2006): Zielsetzung: Ergebnisorientierte Schul- und Unterrichtsentwicklung. In: Böttcher, W./ Holtappels, H.-G./Brohm, M. (Hrsg.): Evaluation im Bildungswesen – Eine Einführung in Grundlagen und Praxisbeispiele. – Weinheim, S. 177–193.

Peek, R./Steffens, U./Köller, O. (2006): Positionspapier des Netzwerks Empiriegestützte Schulentwicklung (EMSE) zu: Zentrale standardisierte Lernstandserhebungen. – 5. EMSE-Tagung, Berlin 08.12.2006.

Peek et al. 2006 = Peek, R./Pallack, A./Dobbelstein, P./Fleischer, J./Leutner, D. (2006): Lernstandserhebungen 2004 in Nordrhein-Westfalen – zentrale Testergebnisse und Perspektiven für die Schul- und Unterrichtsentwicklung. In: Eder, F./Gastager, A./Hofmann, F. (Hrsg.): Qualität durch Standards. – Münster, S. 219–233.

Schmitz, G. S./Schwarzer, R. (2000): Selbstwirksamkeitserwartung von Lehrern: Längsschnittbefunde mit einem neuen Instrument. In: Pädagogische Psychologie, Bd. 14, H. 1, S. 12–25.

Schrader, F.-W./Helmke, A. (2004): Von der Evaluation zur Innovation? Die Rezeptionsstudie WALZER: Ergebnisse der Lehrerbefragung. Empirische Pädagogik, 18 (1), S. 140–161.

Schwarzer, R./Jerusalem, M. (1999): Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen. – Berlin.

Sill, H.-D./Sikora, C. (2007): Leistungserhebungen im Mathematikunterricht – Theoretische und empirische Studien. – Hildesheim.

Tresch, S. (2007): Potenzial Leistungstest. Wie Lehrerinnen und Lehrer Ergebnisrückmeldungen zur Sicherung und Steigerung ihrer Unterrichtsqualität nutzen. – Bern.

Visscher, A. J./Coe, R. (2003): School performance feedback systems: Conceptualisation, Analysis, and Reflection. In: School effectiveness and school improvement, Bd. 14, H. 3, pp. 321–349.

Wall, D. (2000): The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? In: System, Vol. 28, pp. 499–509.