Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 3 (2011), No. 2, 62–93

Uwe Maier, Thorsten Bohl, Marc Kleinknecht & Kerstin Metz

Einflüsse von Merkmalen des Testsystems und Schulkontextfaktoren auf die Akzeptanz und Rezeption von zentralen Testrückmeldungen durch Lehrkräfte

Zusammenfassung

In Deutschland wurden standardbasierte Vergleichsarbeiten eingeführt, um die Qualität im öffentlichen Schulsystem zu sichern und zu verbessern. Die Wirkung dieser Tests hängt jedoch wesentlich von der schulinternen Rezeption und Nutzung der Leistungsrückmeldungen durch Lehrkräfte ab. Ein theoretisches Rahmenmodell legt nahe, dass sowohl Merkmale des Testsystems als auch der Implementation dabei eine Rolle spielen und institutionelle Kontextfaktoren auf Ebene des Schulsystems und Ebene der Schule den Effekt moderieren. In einer quantitativen Vorstudie zu Vergleichsarbeiten in Baden-Württemberg und Thüringen konnte gezeigt werden, dass sich Unterschiede im Testsystem und der Testimplementation auf die Akzeptanz und Nutzung der Testrückmeldungen durch Lehrkräfte auswirken können. Das Ziel dieser Studie war eine weiterführende, vertiefte Analyse der Befunde auf der Basis einer wesentlich größeren Lehrerstichprobe (n = 1777) und unter Kontrolle weiterer Kontextfaktoren auf Ebene der Einzelschule (Lehrerkooperation, Diskussion von Testrückmeldungen in Gremien). In multiplen Regressionsanalysen und einer multivariaten Varianzanalyse erwiesen sich die theoriekonformen Differenzen in der Wahrnehmung der curricularen Validität, der diagnostischen Hinweise und

Prof. Dr. Uwe Maier (corresponding author), Lehrstuhl für Schulpädagogik, Friedrich-Alexander-Universität Erlangen-Nürnberg, Regensburger Straße 160, 90478 Nürnberg, Deutschland
E-Mail: uwe.maier@ewf.uni-erlangen.de

Prof. Dr. Thorsten Bohl, Institut für Erziehungswissenschaft, Abteilung Schulpädagogik, Universität Tübingen, Münzgasse 22–30, 72070 Tübingen, Deutschland
E-Mail: thorsten.bohl@uni-tuebingen.de

Dr. Marc Kleinknecht, Lehrstuhl für Schulpädagogik, TUM School of Education, Schellingstraße 33, 80799 München, Deutschland
E-Mail: marc.kleinknecht@tum.de

Dr. Kerstin Metz, Institut für Sprachen, Pädagogische Hochschule Ludwigsburg, Reuteallee 46, 71634 Ludwigsburg, Deutschland
E-Mail: metz@ph-ludwigsburg.de

der Einschätzung negativer Folgen von Vergleichsarbeiten zwischen den beiden Bundesländern als stabil. Es wird diskutiert, in welchem Maße die Unterschiede im Testsystem und andere institutionelle Kontextfaktoren für diesen Befund verantwortlich gemacht werden können.

Schlagworte
Vergleichsarbeiten; Testbasierte Rechenschaftslegung; Rückmeldungen; Schülerleistungen; Unterrichtsentwicklung; Lehrer

Impact of mandatory testing system and school context factors on teachers' acceptance and usage of school performance feedback data

Abstract
The states of the Federal Republic of Germany implemented educational standards and mandatory testing for controlling and improving the quality of public schooling. The success of this reform highly depends on teachers' acceptance and sensemaking of external feedback data. A theoretical model suggests that features of the testing system and institutional context factors on the system level and the school level moderate teachers' interpretation and usage of school performance feedback. A preliminary cross-country teacher survey revealed differences in teacher acceptance of external testing feedback between two German states with different approaches to mandatory testing. This paper aimed at testing if the country difference remains stable when analyzing a larger sample size (n = 1,777) and controlling for more contextual factors on the school level (test subjects, tracking, teacher cooperation, discussion of performance feedback in faculty meetings). Multiple regression analyses and a multivariate analysis of variance and covariance approved the difference in teachers' perspective on curricular validity, diagnostic usage and negative consequences of mandatory testing between the two German states. The paper eventually discusses to what extent the states' testing system and other institutional context variables account for the differences in teachers' view on performance feedback data.

Keywords
Mandatory testing; Test-based school reform; School performance feedback; Academic achievement; Instructional improvement; Teacher

1. Theoretischer Hintergrund

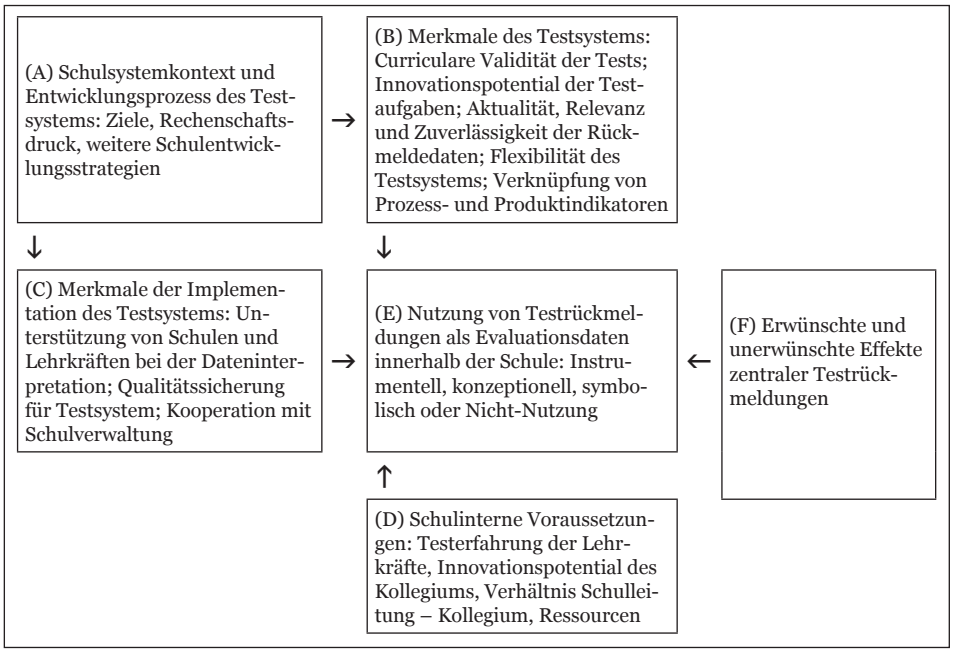
Elemente testbasierter Schulreform wie Bildungsstandards, Vergleichsarbeiten und *large scale assessments* gehören seit gut einem Jahrzehnt zum Standardrepertoire der Reformpolitik im Schulsystem (Baumert, 2001; Klieme, 2004; Peek, Steffens &

Köller, 2006). In grober Anlehnung an testbasierte Reformstrategien im angloamerikanischen Raum verspricht man sich durch die zyklische Messung von grundlegenden Schülerkompetenzen Impulse zur Verbesserung der Bildungsqualität auf allen Systemebenen. Im Gegensatz zum angloamerikanischen Raum verzichtet man in Deutschland jedoch weitgehend auf testbasierten Rechenschaftsdruck durch Veröffentlichung der Leistungsdaten oder andere Sanktionsmechanismen und setzt auf schulinterne Entwicklungsimpulse durch zentrale Testrückmeldungen (Herzog, 2010).

1.1 Ein Modell zur Erklärung der schulinternen Nutzung von externen Testrückmeldungen

Inwiefern dieses Reformversprechen eingelöst werden kann, ist im angloamerikanischen Raum bereits seit Jahrzehnten Thema sowohl einer normativen und oft kontrovers geführten Debatte als auch zahlreicher empirischer Studien. Vor allem die empirische Untersuchung von Effekten testbasierter Schulreformen ist ein komplexes und noch nicht geklärtes Unterfangen (deutschsprachige Zusammenfassung: Maier, 2010). Zur Ordnung der verschiedenen Einflussfaktoren auf die schulinterne Nutzung bzw. die intendierten und nicht intendierten Effekte von externen Schulleistungsdaten eignet sich ein von Visscher und Coe (2003) entwickeltes und von Verhaeghe, Vanhoof, Valcke und Van Petegem (2010) weiterentwickeltes Rahmenmodell (SPFS-model: school performance feedback system model). Dieses Modell unterscheidet zwischen dem Entwicklungsprozess des Testsystems, den zentralen Merkmalen des Testsystems, den Bedingungen der Implementation und den Einzelschulmerkmalen (Abbildung 1). Akzeptanz und Nutzung der Rückmeldedaten durch Lehrkräfte werden dann als wichtige Mediatorvariablen für erwünschte und unerwünschte Nebeneffekte zentraler Tests betrachtet. Damit eignet sich dieses Modell, um eine Vielzahl von Einzelergebnissen der empirischen Forschung im Bereich der Effekte testbasierter Schulreformen zu ordnen.

Abbildung 1: *School performance feedback system*-Rahmenmodell nach Visscher und Coe (2003)



(A) Schulsystemkontext und Entwicklungsprozess des Testsystems (*context-related factors*): Test- und Rückmeldesysteme können unterschiedliche Ziele verfolgen. Im Vordergrund steht natürlich immer die Verbesserung der schulischen Leistung. Es gibt jedoch unterschiedliche Vorstellungen darüber, wie dieses Ziel erreicht werden könnte. Mit Testsystemen kann man beispielsweise die elterliche Schulwahl unterstützen oder man zielt auf die interne Schulentwicklung. Zur Frage, ob externer Druck die schulinterne Nutzung von Testdaten zur Optimierung von Unterricht stimulieren kann, weisen ländervergleichende Studien innerhalb der USA eher auf positive Effekte hin, allerdings nur unter der Voraussetzung, dass testbasierter Rechenschaftsdruck mit weiteren Unterstützungsmaßnahmen gekoppelt wird. Von Bedeutung ist ebenso, dass die mit dem Testsystem verknüpften Ziele mit der Rückmeldekonzeption interagieren. Je nach Zielsetzung sind unterschiedliche Leistungsindikatoren von Bedeutung. Für Schulentwicklungszwecke sollten die Rückmeldungen detailliert sein, um der Komplexität von Schule gerecht werden zu können. Ebenso sollte das *standardisation-flexibility*-Problem beachtet werden. Testsysteme müssen bestimmte Standardindikatoren bereitstellen, die für alle Schulen relevant sind. Es muss jedoch auch noch Raum für individuelle Datenwünsche der Schulen bleiben.

(B) In einem weiteren Variablenbereich werden Merkmale des Testsystems, die sich auf die Nutzung positiv auswirken können, zusammengefasst: Relevant ist vor allem die Qualität der erzeugten Evaluationsdaten. Handelt es sich aus

Sicht der Lehrkräfte um aktuelle, zuverlässige und valide Leistungsdaten? Valide sind Rückmeldedaten dann, wenn sie tatsächlich etwas über die Lernleistung der Schüler und die Unterrichtseffektivität der Lehrkraft bzw. der ganzen Schule aussagen. Geben die Daten des Rückmeldesystems Auskunft über Informationen wie: Trends, Beziehungen zwischen Variablen, z.B. zwischen Leistungsindikatoren einerseits und Prozessindikatoren (Skalen zur Unterrichtsqualität) andererseits? Erlaubt das schulische Leistungsrückmeldesystem eine differentielle Analyse der Informationen? Können die Lehrkräfte bzw. Schulleiter einzelne Subgruppenergebnisse (Schüler mit besonderen Lernschwierigkeiten) analysieren? Werden die Ergebnisberichte auf Wunsch der Schule angepasst?

Die Validität der Testrückmeldungen kann beispielsweise durch die Bereitstellung von fairen Vergleichsdaten oder sog. *value added data* gewährleistet werden. Dabei handelt es sich um längsschnittlich gemessene Leistungszuwächse, d.h. im Testsystem werden die Leistungsstände von Schülern vor und nach einem bestimmten Bildungsabschnitt erfasst. Die in Deutschland praktizierten fairen Vergleiche entsprechen nicht dem *value-added*-Ansatz, weil zwar relevante Prädiktoren für die Schulleistung (z.B. Geschlecht, sozioökonomischer Hintergrund) herausgerechnet werden, es dennoch zu keiner Berücksichtigung des Vorwissens kommt. Ebenso hängt die Validität der Daten von der Passung der Testaufgaben mit dem Lehrplan und den bereits unterrichteten Lerninhalten (implementiertes Curriculum) ab. Die Zuverlässigkeit der rückgemeldeten Daten lässt sich durch eine zusätzliche Kontrolle bei der Testauswertung steigern. Ebenso spielt es eine Rolle, mit welchen statistischen Verfahren die Daten ausgewertet werden (klassische oder probabilistische Testtheorie). Sowohl bei Visscher und Coe (2003), als auch bei Verhaeghe et al. (2010) nicht aufgeführt und thematisiert ist die Bedeutung der Testaufgaben. So ist zu fragen, inwieweit Aufgabenstellungen mit einem fachdidaktischen Innovationspotential aufgenommen wurden. Die *Washback*-Forschung konnte beispielsweise zeigen, dass innovative Testaufgaben (z.B. Aufgaben zur Textproduktion in Fremdsprachtests) von Lehrkräften durchaus aufgegriffen werden und in bestimmten Teilbereichen zu einer Veränderung des zukünftigen Unterrichts führen können (Cheng, 1999; Firestone, Winter & Fitz, 2000; Cheng & Curtis, 2004).

(C) Auch die Merkmale der Implementation des Testsystems spielen eine nicht unbedeutende Rolle für die Erklärung der Nutzung von Rückmeldedaten in den Einzelschulen (*support-related factors*): Werden die Tests und die Leistungsrückmeldungen in weitere, langfristig angelegte Reformstrategien eingebettet oder handelt es sich um isolierte oder sogar kontraproduktive Reformvorhaben? In welchem Umfang werden Schulen und Lehrkräfte bei der Interpretation der Daten sowie der Ableitung und auch Umsetzung von Konsequenzen unterstützt? Werden die Beteiligten (Lehrkräfte, Schulleitungen) in besonderem Maße motiviert, sich das Test- und Rückmeldesystem anzueignen (*ownership*)? Werden die Tests als eine hilfreiche, wertvolle Innovation dargestellt (evtl. um den Lehrerberuf weiter zu professionalisieren)? Werden die Implementation und deren Folgen, d.h. die Wirkung auf Unterricht und die schu-

lische Leistung der Schüler überwacht? Finden laufende Evaluationsstudien statt, um den Prozess der Einführung von zentralen Testsystemen zu evaluieren und optimieren zu können?

Von erfolgversprechenden Modellen zur Implementation zentraler Testsysteme berichteten eine Reihe von Autoren (Yang et al., 1999; Hayes & Rutt, 1999; Saunders, 2000; Wikeley, Stoll & Lodge, 2002; Demie, 2003; Peng, Thomas, Yang & Li, 2006). Ein wichtiges Ergebnis dieser Literatur ist, dass die schulinterne Nutzung externer Leistungsdaten von außen gezielt unterstützt werden muss. Wikeley, Stoll und Lodge (2002) berichten beispielsweise, dass Lehrer durch die Teilnahme an spezifischen Fortbildungen zusehends in die Lage versetzt wurden, externe Leistungsdaten schüler- und schülergruppenspezifisch zu interpretieren. Daraufhin konnten spezifische Programme für diese Schülergruppen entwickelt und durchgeführt werden. Wenn allerdings aufgrund eines schmalen Projektbudgets keine Fortbildungen und Besprechungen stattfanden, konnten Schulen mit ohnehin geringen Veränderungskapazitäten die externen Leistungsinformationen nicht nutzen.

Die vertrauensvolle Zusammenarbeit mit den lokalen Schulbehörden ist eine ebenso wichtige Voraussetzung für eine sinnvolle Dateninterpretation auf Schulebene (Rudd & Davies, 2002; Opfer, Henry & Mashburn, 2008). Als besonders hilfreich stellte sich heraus, wenn lokale Schulbehörden Leistungsindikatoren auswählen, die zuverlässige Aussagen über die Lernerfolge einer Schule zulassen (z.B. langfristige Trends im Bereich der Lesekompetenz), und diese Leistungsindikatoren mit weiteren organisationalen Variablen (z.B. Schülerfehlzeiten, Unterrichtsausfälle, Fördermaßnahmen, etc.) verknüpfen und in lesbarer Form den Schulen zur Verfügung stellen (Demie, 2003; Louis, Febey & Schroeder, 2005). Die Schnittstelle dieser Zusammenarbeit zwischen Schulverwaltung und Einzelschule sind Experten, die sich auf die Dateninterpretation und -nutzung spezialisiert haben (Hayes & Rutt, 1999).

(D) Ein vierter Variablenbereich auf institutioneller Ebene sind die Einstellungen und Fähigkeiten der Lehrkräfte bzw. der Einzelschule als Organisation (*user-related factors*): Die Kapazität einer Schule, sich mit Leistungsrückmeldungen produktiv auseinanderzusetzen, steigt beispielsweise mit der Testerfahrung und dem Vorhandensein von Personen, die eine spezifische Expertise im Umgang mit Testrückmeldungen zur Verfügung stellen können. Studien zeigen, dass Lehrkräfte mit einer gewissen statistischen Expertise bei der Übersetzung und Interpretation von Daten helfen können (Yang et al., 1999; Wikeley, Stoll & Lodge, 2002). Harris, Jamison und Russ (1995) beschreiben in diesem Zusammenhang die Rolle der Fachabteilungsleiter. Ihnen muss es gelingen, fachdidaktisch orientierte Visionen im Kollegium zu bündeln und diese mit den Implikationen externer Leistungsindikatoren zu verknüpfen. Zentrale Testergebnisse werden auf Abteilungsebene vor allem dann systematisch analysiert, wenn sich das Lehrerkollegium bereits über die Notwendigkeit von Reformmaßnahmen verständigt hat. Unterstützend wirkte zudem die schulinterne Belohnung von Lehrern und Abteilungen, wenn aufgrund der Datenrückmeldung unterrichtliche

Innovationen erprobt wurden. Von Vorteil war ebenso, wenn Lehrer die korrigierten Leistungsdaten ihrer Schüler mit weiteren Performanzindikatoren oder qualitativen Beobachtungen in Verbindung bringen konnten (Saunders, 2000).

Auch das generelle Innovationsklima an Schulen ist eine weitere, wichtige Kontextvariable. In Schulen mit geringen Veränderungskapazitäten beispielsweise standen die entwickelten Testrückmeldesysteme recht unverbunden anderen Maßnahmen gegenüber (Wikeley, Stoll & Lodge, 2002). Dagegen kann sich die Erfahrung der Schule mit dem staatlichen Test- und Rückmeldesystem positiv auswirken. Die Intensität der Datennutzung steigt mit der Gesamtdauer einer Schule in einem Rückmeldeprojekt (Saunders, 2000; Louis, Febey & Schroeder, 2005; Yang et al., 1999). Die Nutzung externer Leistungsdaten muss somit als Prozess an den Schulen betrachtet werden.

(E) Diese vier Variablenbereiche A bis D auf institutioneller Ebene wirken sich auf die tatsächliche Rezeption und Nutzung durch Lehrkräfte bzw. Kollegien aus. Verhaeghe et al. (2010) konzeptualisieren die schulinterne Nutzung mit einem sog. *policy-making cycle,* der vergleichbar ist mit dem Zyklenmodell von Helmke und Hosenfeld (2005). Die Dissemination, Wahrnehmung und Interpretation der Befunde wird von verschiedenen Akteuren bzw. Gremien innerhalb der Schule in einem zyklischen Prozess geleistet. Visscher und Coe (2003) betonen allerdings qualitativ unterschiedliche Typen der Nutzung von Rückmeldedaten innerhalb dieses zyklischen Prozesses (in Anlehnung an Rossi, Lipsey & Freeman, 2004):

1. Bei einer instrumentellen Nutzung werden aufgrund der zur Verfügung stehenden Leistungsinformationen konkrete Entscheidungen getroffen.
2. Konzeptuelle Nutzung bedeutet, dass die Rückmeldung eher generell das Denken der Entscheidungsträger beeinflusst.
3. Im Gegensatz dazu sprechen die Autoren von symbolischer Nutzung, wenn die Feedback-Informationen selektiv genutzt werden, um den eigenen, bereits feststehenden Standpunkt argumentativ zu stützen.

(F) Eine instrumentelle bzw. konzeptuelle Nutzung von Testrückmeldungen ist dann die notwendige Voraussetzung für tatsächliche Veränderungen innerhalb der Schule. Im Modell werden erwünschte und unerwünschte Effekte von zentralen Testsystemen in drei nicht näher spezifizierte Bereiche unterteilt:

a) Höhere Leistungserwartungen an die Schüler.
b) Verbesserter Unterricht.
c) Verbesserung der Schulorganisation und der internen Kommunikation. Veränderungen in diesen drei Bereichen können sich auf die Schülerleistungen auswirken.

1.2 Rezeptionsforschung im deutschsprachigen Raum

Mit Hilfe dieses Rahmenmodells können Studien zur Rezeption und Nutzung von Bildungsstandards und Vergleichsarbeiten in Deutschland sortiert und bewertet werden. Dies ist zunächst einmal möglich, weil Visscher und Coe (2003) keine Aussagen machen, die sich nur auf einen bestimmten Typ externer Testsysteme (z.B. *high stakes testing* mit hohem Rechenschaftsdruck) beziehen, sondern allgemein relevante Kontextfaktoren für externe Leistungsrückmeldesysteme identifizieren. Für die Weiterentwicklung und Spezifizierung des Modells wäre es deshalb von Bedeutung, Befunde aus Ländern mit eher geringem, testbasierten Rechenschaftsdruck zu integrieren.

In der deutschsprachigen Literatur finden sich vor allem Evaluations- oder *usability*-Studien zur Rezeption und Nutzung von Vergleichsarbeitsrückmeldungen (z.B. Moser, 2003; Nachtigall, 2005; Nachtigall & Kröhne, 2006; Groß Ophoff, Koch, Hosenfeld & Helmke, 2006; Bonsen, Büchter & Peek, 2006; Hosenfeld & Groß Ophoff, 2007; Tresch, 2007; Sill & Sikora, 2007; Kuper & Hartung, 2007; Maier, 2008a). Diese Studien beschreiben in der Regel die Art und Weise der Nutzung von Testfeedback durch Lehrkräfte und Schulen (Teil E des theoretischen Rahmenmodells). In vielen Studien wird das Zyklenmodell von Helmke und Hosenfeld (2005) oder ein ähnliches Evaluationsmodell (z.B. Tresch, 2007) herangezogen. Die wichtigsten Befunde dieser Studien lassen sich folgendermaßen zusammenfassen:

Die Vergleichsarbeiten werden von Lehrkräften überwiegend als informativ wahrgenommen und die Durchführung sowie die Auswertung bereiten in der Regel kaum Probleme. Im Vordergrund des Interesses aus Lehrersicht stehen allerdings eher der soziale Vergleich (Klasse im Landesmittelwert) und die Einschätzung einzelner Schüler (bzw. Vergleich mit eigener Notengebung). Nur in einem geringeren Maße wird dagegen die Bedeutung der Tests für die Reflexion des eigenen Unterrichts gesehen. Eine offene, schulinterne Diskussion über die Leistungsdaten und Konsequenzen ist eher selten (Nachtigall, 2005; Maier, 2008a). Schulleiter sind vor allem dann skeptisch, wenn die Daten an die Schulverwaltung weitergeleitet werden (Sill & Sikora, 2007). Mittelmäßige oder gute Ergebnisse waren für die Schulleiter zufriedenstellend und können keine weitergehende Reflexion anregen. Sehr schlechte Ergebnisse wurden überwiegend external attribuiert (Unterrichtsausfall, Schüler kannten Aufgabenformate nicht, etc.). Als bedeutsam für die schulinterne Diskussion der Daten kristallisiert sich der subjektiv empfundene Nutzen zentraler Testrückmeldungen heraus (Bonsen, Büchter & Peek, 2006).

Ein ähnliches Muster zeigt sich auch bei Studien zur Rezeption von Bildungsstandards als Voraussetzung für eine testdatenbasierte Schulentwicklung (z.B. Wacker, 2008; Pant et al., 2008). Pant, Vock, Pöhlmann und Köller (2008) legten eine Studie vor, die, ausgehend von einem *concern*-theoretischen Ansatz, die kognitive und emotionale Beschäftigung von Lehrkräften mit standardbasierten Reformen in Deutschland untersucht. Die Resultate dieser Studie weisen darauf hin, dass Lehrkräfte kurz nach der Implementation standardbasierter

Reformen mehrheitlich auf der Suche nach Informationen sind und sich mit möglichen Auswirkungen auf die eigene Person und die kollegiale Kooperation auseinandersetzen. Dagegen kommt es eher kaum zu einer Auseinandersetzung mit den praktischen Konsequenzen der Implementation von Bildungsstandards. Pant et al. (2008) interpretieren diesen Befund dahingehend, dass viele Lehrkräfte standardbasierte Unterrichtsreformen entweder noch gar nicht als solche identifizieren können oder nur als theoretische Innovationsanforderung verstehen.

Über diese deskriptiven Befunde hinaus gibt es nur sehr wenige empirische Studien, in denen Zusammenhänge zwischen Merkmalen des Testsystems oder der Testimplementation und der schulinternen Nutzung untersucht werden (Teile A–C des Rahmenmodells). Es gibt beispielsweise Hinweise, dass sich durch geeignete Unterstützungs- und Kontrollmaßnahmen die Nutzung der Daten für die Schul- und Unterrichtsentwicklung steigern lässt. Im Schweizer Projekt Check 5 mussten die Lehrkräfte Maßnahmen im Anschluss an die Tests schriftlich fixieren. Dies steigerte die unterrichtsbezogene Reflexion der Testergebnisse (Tresch, 2007). Bisher zeigten sich auch fachspezifische Differenzen. Mathematiklehrkräfte bewerten Testrückmeldungen generell positiver und schätzen ihren Nutzen höher ein als Lehrkräfte in Deutsch (vgl. Blum, Drüke-Noe, Leiß, Wiegand & Jordan, 2005; Tresch, 2007; Sill & Sikora, 2007; Maier, 2008a).

Darüber hinaus gibt es eine eher theoretisch bzw. normativ geprägte Diskussion über „gute“ Testaufgaben und Testrückmeldungen, die sowohl messtheoretischen Anforderungen genügen, als auch für Lehrkräfte ein gewisses Innovationspotential entfalten können (Büchter & Leuders, 2005a, 2005b; Blum et al., 2005; Lorenz, 2005; Nachtigall & Kröhne, 2006; Peek & Dobbelstein, 2006; Sill & Sikora, 2007). Diese Diskussion deckt sich mit den in Teil B des Modells von Visscher und Coe (2003) beschriebenen Forderungen nach Validität, Aktualität und Zuverlässigkeit von Rückmeldedaten. Um die Praktikabilität und Relevanz der Daten für Lehrkräfte zu sichern, sollten beispielsweise Kompetenzprofile und Aufgabenlösungshäufigkeiten auf Klassenebene zurückgemeldet werden. Diese Daten sind aufgrund des Aggregationsniveaus hinreichend reliabel und ermöglichen der Lehrkraft eine Gesamtbeurteilung der durch den Unterricht aufgebauten Kompetenzen vor dem Hintergrund der angebotenen sozialen und kriterialen Bezugsnormen. Zentrale Tests können vor allem dann hilfreich sein, wenn sie den Lehrkräften genug Orientierung für den Unterricht geben, z.B. durch gezielte Fehleranalysen und fachdidaktische Aufgabenkommentierungen. Die Rückmeldung der Klassenergebnisse sollte zudem zeitnah erfolgen, damit Lehrkräfte bereits im laufenden Schuljahr auf Defizite reagieren können. Eine empirische Überprüfung, inwiefern diese Kriterien guter Testrückmeldungen für die datenbasierte Schul- und Unterrichtsentwicklung relevant sind, steht jedoch noch aus.

2. Forschungsdesiderat und Hypothesen

Aufgrund der forschungsmethodologischen Anlage erlauben die bisherigen Rezeptionsstudien in Deutschland nur eingeschränkte Aussagen zu empirisch nachweisbaren Effekten zentraler Qualitätsmerkmale von Testsystemen (Teile A–C des Modells) auf die Nutzung dieser Daten in den Schulen. Die Studien in einzelnen Bundesländern sind untereinander nicht vergleichbar, weil sie mit unterschiedlichen Erhebungsinstrumenten arbeiten. Es ist jedoch anzunehmen, dass sich Merkmale des Testsystems (v.a. Aufgaben- und Rückmeldeformate), der Implementation, aber auch schulinterne Voraussetzungen auf institutioneller Ebene auf die Akzeptanz und Nutzung durch Lehrkräfte auswirken. Die dem Bildungsföderalismus geschuldete Variation bundesländerspezifischer Regelungskontexte testbasierter Schulreform wurde deshalb von Maier (2008b) für eine ländervergleichende Studie genutzt. Die Rezeption und Nutzung von zentralen Testrückmeldungen in der Sekundarstufe I wurde in zwei Bundesländern, die sich hinsichtlich Testkonzeption und Vergleichsarbeitsrückmeldungen wesentlich unterscheiden, vergleichend untersucht.

Baden-Württemberg führte verpflichtende Vergleichsarbeiten in den Fächern Deutsch, Mathematik und erste Fremdsprache am Ende der Schuljahre 6 und 8 im Schuljahr 2005/06 ein und nimmt als einziges Bundesland bisher noch nicht an VERA 8 teil. Ziele sind Qualitätssicherung und Feststellung des Lernstands der Schüler. Ab dem Schuljahr 2008/09 wurde der Zeitpunkt der Vergleichsarbeiten auf den Beginn der Schuljahre 7 und 9 gelegt. Das Institut für Schulentwicklung in Stuttgart stellt den Schulen die Testaufgaben zusammen mit selbstauswertenden Excel-Mappen zur Verfügung (http://www.dva-bw.de). Die Lehrkräfte werten die Vergleichsarbeiten direkt nach der Durchführung aus und geben die Punkte in die Selbstauswertungstabelle ein. In dieser Tabelle sind Vergleichswerte aus einer vorher durchgeführten Pilotierungsstudie einprogrammiert. Damit stehen den Lehrkräften in Baden-Württemberg eine Sofortauswertung auf Schüler- und Klassenebene, Angaben in einzelnen Schwerpunktbereichen des Tests sowie ein Vergleich mit den Rohwerten der Pilotierungsstichprobe zur Verfügung. Die Schüler der Pilotierungsstichprobe werden dabei insgesamt in drei Leistungsgruppen eingeteilt. Die Grenzen zwischen den drei Leistungsgruppen werden durch das 25 %-Quartil und das 75 %-Quartil festgelegt. Diese Grenzwerte sind in den Auswertungsmappen hinterlegt und führen nach der Dateneingabe zu einer sofortigen Einteilung der Schüler in drei Leistungsgruppen. Die Schulen erhalten mittlerweile auch eine Auswertungsmappe „Schule“, in der sowohl die einzelnen Klassenergebnisse als auch die Schulergebnisse für die Schulleitungen bzw. die entsprechende Arbeitsgruppe in der Schule aufgeführt werden. Die Benotung der Vergleichsarbeiten war zunächst zwingend und wurde mittlerweile per Erlass untersagt.

Thüringen entwickelte bereits kurz nach Erscheinen der ersten PISA-Studie eigenständige Vergleichsarbeiten, sog. Kompetenztests, für die Fächer Deutsch,

Mathematik und Englisch gegen Ende der Klassenstufen 6 und 8 (http://www.kompetenztest.de). Vergleichbar mit dem Vorgehen in Baden-Württemberg ist die Testdurchführung. Die Lehrkräfte erhalten fertige Testmappen, führen die Kompetenztests bzw. VERA 8 an einem zentral festgelegten Termin durch und werten die Tests nach bestimmten Vorgaben aus. Im Gegensatz zum Vorgehen in Baden-Württemberg geben die Lehrkräfte in Thüringen die Testpunkte für jede Aufgabe und jeden Schüler in eine Online-Datenmaske ein. Die Auswertungen werden von einer Arbeitsgruppe an der Universität Jena vorgenommen. Die Übereinstimmungswerte stichprobenartiger Zweitkorrekturen mit den Erstkorrekturen werden im Landesbericht veröffentlicht. Die Lehrkräfte und Schulen erhalten Rückmeldedaten in insgesamt drei Berichtswellen (Sofortbericht nach ca. 2 Wochen, ausführlicher Klassenbericht, Schulbericht).

Kernstück der Vergleiche sind sog. „faire Vergleiche“, in denen unterrichtsunabhängige Prädiktoren der Schulleistung berücksichtigt werden (z.B. Geschlecht, Migrationshintergrund, sozioökonomischer Hintergrund, Sprachprobleme). Im Vergleich zu Baden-Württemberg stehen den Lehrkräften und Schulen in Thüringen somit landesweite Vergleichswerte zur Verfügung, die die Lernvoraussetzungen der Schüler statistisch berücksichtigen. In Baden-Württemberg wird den Lehrkräften geraten, den Vergleich der Testwerte mit den Rohwerten der Pilotierungsstichprobe vor dem Hintergrund des eigenen Wissens über die Lernvoraussetzungen der Schüler selbst zu bewerten. Die Benotung der Kompetenztests wird den einzelnen Schulen in Thüringen freigestellt. Selbstevaluation und Schulentwicklung werden in Thüringen zudem durch eine Reihe weiterer Projekte unterstützt. Mit dem ebenfalls zentral entwickelten Instrument „Schüler als Experten für Unterricht (SEfU)“ können Schulen eine interne Schülerbefragung durchführen und mit Landesdaten vergleichen. Die Einführung neuer Instrumente der Qualitätssicherung wurde zudem mit einem Programm zur Förderung der Schulautonomie verbunden (Projekt EVAS: Eigenverantwortliche Schule). Auch im Bereich der Schülerbeurteilung wurden Innovationen erprobt. In Thüringen werden die herkömmlichen Zeugnisse auch in der Sekundarstufe I mittlerweile durch sog. Kompetenzbögen mit Angaben zum Lern- und Arbeitsverhalten der Schüler ergänzt. Diese Kompetenzbögen sollen überwiegend als Diagnose- und Beratungsinstrument genutzt werden.

Damit unterscheiden sich die beiden Bundesländer vor allem hinsichtlich der Implementation des zentralen Testsystems und der Rückmeldung landesweiter Vergleichsdaten. In Thüringen ist erkennbar, dass die Einführung zentraler Tests mit weiteren Schritten sowohl zur Sicherung der Qualität als auch zur Erhöhung der Schulautonomie einhergeht. Bei der Einführung zentraler Tests als Qualitätssicherungsinstrument wurde darüber hinaus besonderer Wert auf faire, adjustierte Landesvergleiche gelegt. Damit werden in Thüringen zwei zentrale Forderungen des theoretischen Rahmenmodells realisiert. Allerdings gibt es vor dem Hintergrund des Rahmenmodells auch große Übereinstimmungen. Es wird in einem jährlichen Rhythmus getestet. Schulen und Lehrkräfte haben keine Mitwirkungsmöglichkeiten bei der Auswahl der zu testenden Inhalte. Ebenso fol-

gen in beiden Bundesländern praktisch keine externen Konsequenzen. Es liegt ganz in der Verantwortung der einzelnen Schule, sinnvoll mit den Daten umzugehen.

In einer quantitativen Lehrerbefragung wurde der Vermutung nachgegangen, dass sich diese Differenzen im Rückmeldesystem und der Implementation von Testdaten auf die Rezeption und Nutzung auswirken (Maier, 2008b). Die Studie zeigte, dass Lehrkräfte in Thüringen die Kompetenztests hinsichtlich Lehrplanvalidität, diagnostischer Hinweise, Hinweisen auf Unterrichtsveränderungen und einer wahrgenommenen Belastung insgesamt positiver bewerteten. Die Lehrkräfte in Baden-Württemberg schätzten dagegen den Nutzen der Rückmeldungen für die eigene Notengebung in dem getesteten Fach höher ein, weil die Vergleichsarbeiten zum Zeitpunkt der Studie noch benotet wurden. In Thüringen wurde den Schulen die Benotung der Tests freigestellt. Um die Stabilität der Länderunterschiede hinsichtlich der Bewertung von Vergleichsarbeiten zu testen, wurden multiple Regressionsanalysen mit weiteren möglichen Einflussvariablen berechnet (Schulform, Schulgröße, Klassengröße, Lehrerselbstwirksamkeitserwartung, Schulleitung). Vor allem das getestete Schulfach und die Klassengröße erwiesen sich als Prädiktoren für die Rezeption und Nutzung der Testdaten. Es gab eine höhere Einschätzung der Belastung durch zentrale Tests in größeren Klassen und eine bessere Bewertung der curricularen Validität bzw. der Nutzung in Mathematik im Vergleich zu den Tests im Fach Deutsch. Deutliche Länderunterschiede gab es auch bei der schulinternen Diskussion der Testrückmeldungen. In Thüringen wurden die Kompetenztestrückmeldungen häufiger und systematischer in den Fach- und Gesamtlehrerkonferenzen diskutiert.

Diese Studie ließ jedoch auch Fragen offen. Außer Schulform und den zwei getesteten Schulfächern (Deutsch, Mathematik) wurden keine weiteren Kontextfaktoren auf institutioneller Ebene erfasst. Vor dem Hintergrund des Rahmenmodells ist jedoch anzunehmen, dass sich weitere Faktoren auf institutioneller Ebene, wie z.B. die kollegiale Diskussion von Testrückmeldungen oder die Einstellung der Lehrkräfte zu Innovationen auf die Rezeption der Testrückmeldungen auswirken. In der Vorgängerstudie wurde die kollegiale Diskussion der Testrückmeldungen jedoch als weitere abhängige Variable behandelt. Die theoretischen Modelle gehen jedoch davon aus, dass es sich um eine Mediatorvariable handelt, die den Effekt des Testsystems auf die Rezeption und Nutzung durch einzelne Lehrkräfte beeinflusst. Ebenso wurde in dieser Studie nur der Unterschied in der Rezeption zwischen Tests in Deutsch und Mathematik untersucht. Unklar blieb, ob sich die Befunde auf Vergleichsarbeiten in den Fremdsprachen übertragen lassen.

Ein weiteres Problem lag in der Stichprobengröße. Insgesamt konnte für diese Studie auf eine Lehrerstichprobe von $n = 1136$ zurückgegriffen werden. Jedoch verringerte sich aufgrund fehlender Werte bei den relevanten Kovariaten die Größe der Stichprobe für die Regressionsanalysen auf $n = 673$. Aus diesem Grund wurden in dieser Studie multiple Regressionsanalysen gerechnet, um den Einfluss der Kontextvariablen zu berücksichtigen. Aufgrund der geringen Fallzahlen wurden zudem nur die Fächer Deutsch und Mathematik miteinander verglichen. Um die Annahme, dass sich das bundeslandspezifische Testsystem auf die Rezeption

und Nutzung der Testdaten unter Kontrolle möglicher Kontextvariablen auswirkt, zu testen, sollten jedoch multivariate, mehrfaktorielle Varianzanalysen mit den Faktoren Land (bzw. Testsystem), Fach und Schulform gerechnet werden. Dies war in dieser Vorstudie aufgrund der geringen Fallzahlen in den Untergruppen (Schulfach, Schulform) nicht möglich.

Ziel dieser weiterführenden Studie war es deshalb, mit einer erweiterten Stichprobe zu prüfen, ob die Länderdifferenzen zwischen Lehrkräften in Thüringen und Baden-Württemberg hinsichtlich der Bewertung von Vergleichsarbeiten stabil bleiben, auch wenn ein weiteres getestetes Fach (Fremdsprachen) und weitere, möglicherweise relevante Kontextfaktoren wie kollegiale Kooperation und Umfang der Diskussion von Testrückmeldungen in Gremien kontrolliert werden. Ebenso sollte der Einfluss der schulinternen Diskussionskultur auf das Rezeptionsverhalten einzelner Lehrkräfte geklärt werden. Aufgrund des theoretischen Rahmenmodells und der Ergebnisse der vorausgehenden Studie wurden folgende Hypothesen geprüft:

1. Das Ausmaß der kollegialen Kooperation in Schulen wirkt sich positiv auf die Bewertung der Rezeption und Nutzung von Vergleichsarbeitsergebnissen durch einzelne Lehrkräfte aus.
2. Diskussionen über Tests und Testrückmeldungen in schulischen Gremien wirken sich positiv auf die Bewertung der Rezeption und Nutzung von Vergleichsarbeitsergebnissen durch einzelne Lehrkräfte aus.
3. Die Bewertung der Rezeption und Nutzung von Vergleichsarbeitsergebnissen durch Lehrkräfte ist in Thüringen günstiger als in Baden-Württemberg, auch wenn Kontextfaktoren auf Lehrerebene, Schulebene sowie Schulfach und Schulart kontrolliert werden.

Bei der dritten Hypothese war mit Wechselwirkungseffekten zu rechnen. Bei bisherigen Studien zeigten sich immer wieder deutliche Schulart- und Fachunterschiede in der Rezeption und Nutzung von Testrückmeldungen. D.h. die hypothetisch angenommene Differenz in der Rezeption zwischen den beiden Ländern könnte in Abhängigkeit von Schulfach und Schulart durchaus variieren.

3. Methodische Vorgehensweise

Durch zusätzliche Datenerhebungen in den Jahren 2009 und 2010 zur Rezeption und Nutzung von Vergleichsarbeiten in den Bundesländern Baden-Württemberg und Thüringen war es möglich, diese weiterführenden Analysen durchzuführen. Es handelte sich nicht nur um eine Replikationsstudie, sondern um eine weiterführende Studie, weil durch die größere Datenbasis weitere Fächervergleiche möglich wurden. Ebenso gestattete es die größere Stichprobe, Hypothesen sowohl mit multiplen Regressionsanalysen als auch mit mehrfaktoriellen Varianzanalysen zu prüfen.

3.1 Stichprobe

Die vorliegenden Daten stammen aus insgesamt vier Erhebungswellen. In einer ersten Erhebungswelle im Jahr 2006 wurden baden-württembergische Lehrkräfte aller Schularten in der Sekundarstufe (HS, RS, GY) zu den neu eingeführten Vergleichsarbeiten befragt (Maier, 2008a). Hierfür wurde per Zufall ca. ein Viertel der Schulen in Baden-Württemberg ausgewählt und postalisch gebeten, an der Studie teilzunehmen. Im darauf folgenden Jahr 2007 wurde mit einem erweiterten Instrumentarium eine erste ländervergleichende Befragung an Hauptschulen, Realschulen und Gymnasien in Baden-Württemberg sowie an Regelschulen und Gymnasien in Thüringen durchgeführt. Pro Bundesland wurden ca. 50 % der Schulen zufällig ausgewählt und wiederum postalisch um Teilnahme gebeten. Die Daten dieser zweiten Erhebungswelle bildeten die Grundlage für die Vorgängerstudie (Maier, 2008b). Um die Effekte speziell in einer Schulform besser nachzeichnen zu können, wurden im Jahr 2009 Lehrkräfte an 50 zufällig ausgewählten Gymnasien in Baden-Württemberg und Thüringen über die Schulleitungen angeschrieben und gebeten, ein abermals erweitertes Fragebogeninstrumentarium auszufüllen und zurückzusenden. Bei der Stichprobenauswahl wurde jeweils darauf geachtet, dass bereits befragte Schulen nicht noch einmal angeschrieben wurden. Da der Rücklauf aufgrund der durchschnittlichen Schulgröße in Thüringen insgesamt geringer war als in Baden-Württemberg, wurden im Jahr 2010 noch einmal 50 weitere Thüringer Gymnasien angeschrieben.

Tabelle 1: Stichprobe nach Schulart und Bundesland

		Schulart		Gesamt
		Hauptschule, Realschule (BW) Regelschule (TH)	Gymnasium	
Bundesland	BW	729	564	1293
		56.4 %	43.6 %	100.0 %
	TH	209	275	484
		43.2 %	56.8 %	100.0 %
Gesamt		938	839	1777
		52.8 %	47.2 %	100.0 %

Damit ergab sich eine nach Bundesland und Schulart relativ gut ausbalancierte Gesamtstichprobe (Tabelle 1). Insgesamt konnten für die weiterführende Analyse Fragebögen von 1293 Lehrkräften in Baden-Württemberg und 484 Lehrkräften in Thüringen ausgewertet werden. Dies entsprach in etwa dem Größenunterschied zwischen den Schulsystemen beider Länder. Da es in Thüringen mittlerweile ein zweigliedriges Schulsystem gibt, wurde in den Varianzanalysen die Variable Schulart dichotomisiert: Gymnasiallehrkraft vs. keine Gymnasiallehrkraft (Regelschullehrkraft in Thüringen; Haupt- oder Realschullehrkraft in Baden-Württem-

berg). Tabelle 2 zeigt die Verteilung der Fragebögen nach Bundesland und Fach, in dem eine Vergleichsarbeit geschrieben wurde.

Tabelle 2: Stichprobe nach getesteten Fächern und Bundesland

		Getestete Schulfächer			Gesamt
		Deutsch	Mathematik	Fremdsprache	
Bundesland	Baden-Württemberg	530	611	89	1230
		43.1%	49.7%	7.2%	100.0%
	Thüringen	119	237	122	478
		24.9%	49.6%	25.5%	100.0%
Gesamt		649	848	211	1708
		38.0%	49.6%	12.4%	100.0%

Insgesamt liegen 69 Fragebögen ohne Angabe eines konkreten Faches oder mit Mehrfachnennungen vor, d.h. es ist unklar, auf welche Vergleichsarbeit die Lehrkräfte ihre Aussagen genau beziehen. Für die Stichprobe in Baden-Württemberg war der prozentuale Anteil der Lehrerfragebögen, die sich auf Vergleichsarbeiten in der ersten Fremdsprache (in der Regel Englisch) beziehen, wesentlich geringer als in Thüringen. Dies liegt vor allem an den unterschiedlichen Regelungen. Während für die Schulen in Thüringen der Kompetenztest in Mathematik und ein weiterer Kompetenztest in Deutsch oder Englisch verpflichtend waren, konnten die Schulen in Baden-Württemberg auf die Vergleichsarbeit in der Fremdsprache verzichten und mussten dafür Tests in Deutsch und Mathematik verpflichtend durchführen.

3.2 Instrumente

Ausgangspunkt war ein selbst entwickelter Fragebogen zu einzelnen Aspekten der Rezeption und Nutzung zentraler Testrückmeldungen für Diagnose, Notengebung und Unterrichtsentwicklung. Lediglich für die allgemeine Akzeptanz zentraler Tests konnte eine bereits entwickelte Skala von Ditton und Merz (2000) genutzt werden. Für jede weitere Befragungswelle wurde das Fragebogeninstrument um weitere Teilaspekte sowie Kontextfaktoren erweitert. Ein Pool von insgesamt 15 zentralen Items zur Erfassung der subjektiven Einschätzung von Vergleichsarbeiten blieb jedoch in allen vier Befragungswellen gleich. Mit einer Hauptkomponentenanalyse mit Varimax-Rotation, die über alle 15 Items gerechnet wurde, konnten insgesamt vier Faktoren extrahiert werden (siehe Anhang). Diese vier Faktoren erklären 66.5 % der Gesamtvarianz und sind Grundlage für die Bildung von vier Skalen mit jeweils 3 bis 6 Items (siehe Anhang):

- Lehrplanvalidität (3 Items; Cronbachs alpha = .80; n = 1702): z.B. „Testaufgaben decken die Lernbereiche des Bildungsplanes ab.“

- Diagnostischer Nutzen des Tests (6 Items; Cronbachs alpha = .88; n = 1701): z.B. „Die Tests geben zusätzliche diagnostische Hinweise."
- Nutzen des Tests für die Notengebung (3 Items; Cronbachs alpha = .71; n = 1740): z.B. „Die Tests regen zum Nachdenken über eigene Bewertungsmaßstäbe an."
- Negative Folgen des Tests (3 Items, Cronbachs alpha = .74; n = 1731): z.B. „Der Test übt zusätzlichen Druck auf Schulen und Lehrer aus."

Diese vier Skalen decken bei weitem nicht alle relevanten Aspekte zur Bewertung von Vergleichsarbeiten und Testrückmeldungen ab. Weitere Bewertungskriterien, wie z.B. das innovative Potential von Testaufgaben, wären denkbar. Dennoch eignen sich diese vier Skalen zu einer Abschätzung des Nutzens von Vergleichsarbeiten aus Lehrersicht. Die von den Lehrkräften wahrgenommene Lehrplanvalidität (in der Regel die Kompatibilität mit dem implementierten Curriculum bzw. dem Schulbuch) ist eine zentrale Voraussetzung, dass die Ergebnisse überhaupt ernst genommen werden. Weiterhin ist das zur Verfügung stellen von objektiver Diagnoseinformation auf Schüler- und Klassenebene ein zentrales Ziel von Vergleichsarbeiten. Die Skala „negative Konsequenzen des Tests" kann als „Stimmungsindikator" für den Grad der allgemeinen Ablehnung zentraler Tests gewertet werden.

Für alle vier Erhebungswellen liegen Daten zu möglichen Kontextbedingungen auf Lehrerebene und institutioneller Ebene vor. Auf institutioneller Ebene werden neben der Länderdifferenz (als Omnibusvariable für die Elaboriertheit der Rückmeldedaten und den Grad an Einbindung zentraler Tests in weitere Schulentwicklungsstrategien) das getestete Fach (Deutsch, Mathematik oder Fremdsprache), die Schulart (Gymnasium vs. kein Gymnasium), die Schulgröße und die Größe der Klasse, in der die Vergleichsarbeit durchgeführt wurde, berücksichtigt. Das Ausmaß der Kooperation im Kollegium wurde mit einer Skala von Ditton, Arnoldt und Bornemann (2002) erfasst (5 Items; Cronbachs alpha = .79; n = 1376: Beispielitem: „An unserer Schule führen Lehrer oft gemeinsame Projekte durch").

Die Lehrkräfte wurden zudem gefragt, in welchen Gremien bzw. mit welchen Personengruppen sie die Testrückmeldungen bisher wie intensiv diskutiert hatten. Folgende Alternativen standen hierfür zur Auswahl: Ausgewählten Kollegen, Klassenkonferenz, Jahrgangsstufenkonferenz, Fachkonferenz, Gesamtlehrerkonferenz, Eltern dieser Klasse. Die Lehrkräfte sollten für jede Personengruppe bzw. für jedes Gremium einschätzen, ob dort die Vergleichsarbeitsrückmeldungen nicht (0), informell (1) oder systematisch (2) diskutiert wurden. Wenn Gremien nicht existieren (z.B. Jahrgangsstufenkonferenzen in kleinen Hauptschulen) konnte „Gremium existiert nicht" angekreuzt werden. Diese Aussage wurde ebenfalls mit „0" codiert.

Auf Lehrerebene liegen Daten zu Geschlecht, Alter und der Lehrerselbstwirksamkeitserwartung (Schwarzer & Jerusalem, 1999; Schmitz & Schwarzer, 2000; 9 Items; Cronbachs alpha = .78; n = 1562) vor. Ebenso wurde erfasst, ob die befragte Person Mitglied der Schulleitung (Schulleiter oder stellv. Schulleiter) ist oder nicht. Keine Angabe in diesem Feld wurde als 0 kodiert, d.h. es wurde an-

genommen, dass diese Probanden nicht Mitglied der Schulleitung in ihrer Schule sind.

4. Ergebnisse

4.1 Deskriptive Ergebnisse

Tabelle 3 zeigt die Häufigkeitsverteilungen der nominalen Variablen und die Durchschnittswerte der metrischen Variablen. Ebenso werden die unterschiedlichen Ausprägungen in den beiden Bundesländern dargestellt und auf Signifikanz geprüft.

In der Thüringer Stichprobe finden sich ein höherer Anteil an Lehrerinnen, im Schnitt um 2 Jahre ältere Lehrkräfte und wesentlich weniger Lehrkräfte mit Vollzeitbeschäftigung wieder. Ursache hierfür sind strukturelle Differenzen zwischen den Schulsystemen beider Bundesländer. In Thüringen arbeiten aufgrund der Beschäftigungssituation in der ehemaligen DDR mehr Lehrerinnen. Die hohe Zahl der Teilzeitstellen korrespondiert mit der Entwicklung nach der Einheit in den 1990er Jahren, als aufgrund des Bevölkerungsrückgangs in Thüringen Lehrkräfte entlassen bzw. Vollzeitbeschäftigungsverhältnisse in Teilzeitstellen umgewandelt wurden.

In der Stichprobe unterscheiden sich die beiden Länder hinsichtlich der prozentualen Anteile der Lehrkräfte, die zugleich Schulleiter oder stellvertretender Schulleiter sind, ebenfalls signifikant voneinander. Auch die kleineren Klassen und die kleineren Schulen der befragten Lehrkräfte in Thüringen korrespondieren mit den schulstrukturellen Gegebenheiten. Ein überwiegender Teil der Schulen in Thüringen liegt in sehr ländlichen Regionen und kleineren Städten. Die Werte für kollegiale Kooperation und Lehrerselbstwirksamkeitserwartung sind leicht günstiger für die Thüringer Lehrkräfte. Dies könnte womöglich mit den kleineren Schulen und Klassen zusammenhängen, die für eine Kooperation eventuell günstigere Ausgangsvoraussetzungen darstellen.

Tabelle 3: Mittelwerte und prozentuale Anteile aller Variablen nach Bundesland

		Gesamt	BW	THÜR	Signifikanztest für Differenz		
Unabhängige Variablen (nominal)	Kategorien	Prozentuale Häufigkeiten			Chi-Quadrat	*df*	Signifikanzniveau
Geschlecht	Lehrerin	65.9%	59.7%	81.7%	96.90	1	$p<.001$
Mitglied Schulleitung	Ja	6.8%	6.0%	8.7%	10.80	1	$p<.01$
Beschäftigungsumfang	Vollzeit	52.0%	59.0%	33.3%	383.37	1	$p<.001$
Unabhängige Variablen (metrisch)	Range Min./Max.	Mittelwerte			*T*-Wert	*df*	Signifikanzniveau
Alter (Jahre)	24–64	45.8	45.1	47.9	-4.72	1679	$p<.001$
Schulgröße (Anzahl 6. Klassen)	1–8	2.8	3.0	2.4	8.42	1724	$p<.001$
Klassengröße	5–35	24.1	25.3	20.6	15.41	1709	$p<.001$
Kollegiale Kooperation	1–5	3.33	3.28	3.43	-3.87	1433	$p<.001$
Lehrerselbstwirksamkeitserw.	1–4	3.00	2.95	3.02	-3.14	1724	$p<.01$
Diskussion mit ausgewählten Kollegen	0–2	1.01	1.00	1.03	-.82	1764	n.s.
Diskussion in Klassenkonferenz	0–2	0.55	0.48	0.73	-7.04	1758	$p<.001$
Diskussion in Jahrgangsstufenkonf.	0–2	0.37	0.33	0.47	-3.95	1758	$p<.001$
Diskussion in Fachkonferenz	0–2	1.02	0.87	1.41	-12.32	1767	$p<.001$
Diskussion in Gesamtlehrerkonferenz	0–2	0.36	0.27	0.61	-10.43	1761	$p<.001$
Abhängige Variablen	Range Min./Max.	Mittelwerte			*T*-Wert	*df*	Signifikanzniveau
Lehrplanvalidität	1–5	3.12	2.97	3.52	-10.88	1749	$p<.001$
Diagnostischer Nutzen	1–5	2.84	2.70	3.22	-9.93	1771	$p<.001$
Nutzen für Notengebung	1–5	2.94	2.97	2.87	1.87	1768	n.s.
Negative Folgen des Tests	1–5	2.49	2.62	2.14	9.09	1752	$p<.001$

Anmerkung. Die metrischen Variablen (ab kollegiale Kooperation) sind so gepolt, dass geringe Werte geringe Ausprägungen und hohe Werte hohe Ausprägungen bedeuten.

Die befragten Lehrkräfte besprechen die Testrückmeldungen am ehesten mit ausgewählten Kollegen oder in der Fachkonferenz. In der Gesamtlehrerkonferenz und der Jahrgangsstufenkonferenz spielen sie so gut wie keine Rolle. Es gibt jedoch deutliche Unterschiede zwischen den Bundesländern. In der Thüringer Stichprobe werden deutlich intensivere Diskussionen in den Klassen-, Fach- und Gesamtlehrerkonferenzen berichtet. Die Länderdifferenzen in der Einschätzung der Lehrplanvalidität, der negativen Folgen sowie des diagnostischen Nutzens bestätigen zunächst einmal die Ergebnisse der Vorgängerstudie. Der Nutzen der zentralen Tests für die Notengebung wird dagegen in beiden Bundesländern gleich hoch eingeschätzt. In der vorangehenden Studie gab es hier noch eine höhere Bewertung seitens der Lehrkräfte in Baden-Württemberg. Dies lässt sich mit

der ab dem Schuljahr 2008/09 in Baden-Württemberg geltenden Regelung, dass Vergleichsarbeiten nicht mehr benotet werden dürfen, erklären.

4.2 Relevanz der Kontextfaktoren auf Schulebene (Hypothesen 1 und 2)

Zunächst wird mit explorativen, multiplen Regressionsanalysen geprüft, wie stark die beiden Kontextvariablen auf Schulebene, kollegiale Kooperation und Diskussion der Testrückmeldungen in Gremien, mit den einzelnen abhängigen Variablen (Rezeption) zusammenhängen. In einem ersten Schritt wurden zunächst die Effekte der kollegialen Kooperation (Hypothese 1) und der Diskussion von Testrückmeldungen in Gremien (Hypothese 2) ohne weitere Prädiktorvariablen auf die Kriteriumsvariablen geprüft (Tabelle 4). Die multiplen Regressionsanalysen zeigen die erwarteten Zusammenhänge. Bis auf die Kriteriumsvariable „Nutzen für die Notengebung" hängt die von den Lehrkräften wahrgenommene kollegiale Kooperation an einer Schule positiv mit der Rezeption und Nutzung von Testrückmeldungen zusammen (bzw. negativ mit den negativen Folgen). Auch die Diskussion der Tests und Testrückmeldungen in den Klassen- und Fachkonferenzen korreliert positiv mit den abhängigen Variablen. Besonders enge Zusammenhänge zeigen sich zwischen Diskussionen in Fachkonferenzen und der Bewertung diagnostischer Hinweise sowie zwischen Diskussionen in Klassenkonferenzen und Hinweisen für die Notengebung.

Tabelle 4: Zusammenhänge zwischen Schulkontext und Rezeption bzw. Nutzung der Tests – Multiple Regressionsanalysen nach dem Einschlussverfahren (Standardisierte Beta-Koeffizienten)

Prädiktorvariablen	Kriteriumsvariablen			
	Lehrplan-validität	Hinweise Diagnostik	Hinweise Noten	Negative Folgen
Kollegiale Kooperation	.09 **	.09 **		-.12 ***
Diskussion. mit ausgewählten Kollegen			.06 *	
Diskussion in der Klassenkonferenz		.08 *	.13 ***	-.10 **
Diskussion in der Jahrgangsstufenkonferenz				
Diskussion in den entsprechenden Fachkonferenzen	.03 **	.17 ***		
Diskussion in der Gesamtlehrerkonferenz				
n	1387	1407	1406	1392
R^2	.022	.063	.021	.028
Korrigiertes R^2	.018	.059	.017	.024
F-Wert	5.14 ***	15.59 ***	5.01 ***	6.68 ***

Anmerkung. Es werden nur signifikante Koeffizienten berichtet.
* $p < .05$, ** $p < .01$, *** $p < .001$.

In einem zweiten Schritt wird nun geprüft, ob diese Zusammenhänge stabil bleiben, wenn weitere, für die Testnutzung möglicherweise relevante Prädiktorvariablen (Alter, Beschäftigungsumfang, Schulleitung, Schulgröße, Klassengröße, Lehrerselbstwirksamkeitserwartung, Fächer, Bundesland) kontrolliert werden. Die nominale Variable, ob eine Lehrkraft Mitglied der Schulleitung ist oder nicht, wird hierzu in den Regressionsanalysen als dichotome Variable behandelt. Die Variable „getestetes Schulfach" wird dummy-rekodiert. Bei diesem Verfahren wird jede der Merkmalskategorien der Originalvariable durch jeweils eine Indikatorvariable repräsentiert, d.h. die Variable „getestetes Fach" wurde in drei dichotome Variablen „Test in Mathematik" (ja/nein), „Test in Deutsch" (ja/nein), „Test in Fremdsprache" (ja/nein) überführt. Die unabhängigen Variablen „Schulform" – Gymnasium (ja/nein) – und „Bundesland" (Baden-Württemberg/Thüringen) liegen ebenfalls bereits als dichotome Variablen vor.

Tabelle 5 zeigt die standardisierten Beta-Koeffizienten der erweiterten, multiplen Regressionsanalysen. Zunächst einmal zeigen sich die bereits in den Vorgängerstudien nachgewiesenen Länderunterschiede. Der Hypothese 3 wird jedoch in einem weiteren Analyseschritt nachgegangen. Auffallend ist, dass sich keine Schulartenunterschiede mehr zeigen. Fachunterschiede ergeben sich nur noch für die Einschätzung der negativen Folgen. Durchweg signifikant korreliert die Klassengröße mit den abhängigen Variablen. Hier könnte sich vor allem die höhere Korrekturbelastung negativ auf die Einschätzung auswirken. Die Tatsache, ob eine befragte Lehrkraft Mitglied der Schulleitung ist oder nicht, wirkt sich lediglich auf die Einschätzung der negativen Folgen aus. Schulleiter bzw. deren Stellvertreter sehen weniger Probleme mit zentralen Tests. An größeren Schulen wird der Nutzen der Testrückmeldungen für eine Adjustierung der Notengebung geringer bewertet. Dieses Ergebnis ist durchaus plausibel, weil Lehrkräfte ihre Leistungsmaßstäbe in einem größeren Kollegenkreis vergleichen können.

Tabelle 5: Zusammenhänge zwischen Kontextfaktoren auf Schul-, Lehrer- bzw. Fachebene und Rezeption bzw. Nutzung der Tests – Multiple Regressionsanalysen nach dem Einschlussverfahren (Standardisierte Beta-Koeffizienten)

Prädiktorvariablen	Kriteriumsvariablen			
	Lehrplan-validität	Hinweise Diagnostik	Hinweise Noten	Negative Folgen
Bundesland (1=BW/2=TH)	.21 ***	.13 ***	-.14 ***	-.13 ***
Gymnasium (0=nein/1=ja)				
Test in Deutsch (0=nein/1=ja)				
Test in Mathematik (0=nein/1=ja)				-.23 *
Test in Fremdsprache (0=nein/1=ja)				-.18 **
Schulgröße			-.14 ***	
Klassengröße	-.07 *	-.14 ***	-.10 **	.15 ***
Mitglied der Schulleitung? (0=nein/1=ja)				-.09 ***
Alter (Jahre)				
Beschäftigungsumfang (Teilzeit/Vollzeit)				
Lehrerselbstwirksamkeitserwartung		.06 *		
Kollegiale Kooperation				
Diskussion mit ausgewählten Kollegen				
Diskussion in der Klassenkonferenz			.10 **	
Diskussion in der Jahrgangsstufenkonferenz				
Diskussion in den entsprechenden Fachkonferenzen		.12 ***		
Diskussion in der Gesamtlehrerkonferenz				
n	1219	1229	1228	1224
R^2	.093	.133	.059	.114
Korrigiertes R^2	.080	.120	.046	.101
F-Wert	7.23 ***	10.82 ***	4.49 ***	9.09 ***

Anmerkung. Es werden nur signifikante Koeffizienten berichtet.
* $p < .05$, ** $p < .01$, *** $p < .001$.

Durch die Hinzunahme weiterer Prädiktorvariablen verschwinden somit die Zusammenhänge zwischen kollegialer Kooperation und der Rezeption und Nutzung von Testrückmeldungen komplett. Die Annahme, dass eine positive kollegiale Zusammenarbeit an einer Schule für sich genommen dazu beiträgt, das Interesse bei Lehrkräften an der Nutzung von Vergleichsarbeitsrückmeldungen zu erhöhen, muss damit relativiert werden (Hypothese 1). Die Effekte der Diskussion von Testrückmeldungen in Gremien auf die Rezeption und Nutzung durch Lehrkräfte reduzieren sich ebenfalls deutlich. Hypothese 2 muss differenziert bewertet werden. Die hier vorliegenden Daten legen die Aussage nahe, dass Lehrkräfte die diagnostischen Hinweise in Testrückmeldungen dann positiver einschätzen, wenn

die Daten auch in den Fachkonferenzen diskutiert wurden. Analog gibt es einen stabilen Zusammenhang zwischen der Diskussion von Testrückmeldungen in den Klassenkonferenzen und der Bewertung des Nutzens von Vergleichsarbeiten für die Notengebung.

4.3 Relevanz des Testsystems (Hypothese 3)

Es wird angenommen, dass die abhängigen Variablen aufgrund der gemeinsamen Thematik (Rezeption und Nutzung von Testrückmeldungen) miteinander korrelieren. Die Korrelationsanalyse bestätigt diese Annahme (Tabelle 6). Die Zusammenhänge zwischen den abhängigen Variablen sind signifikant und in ihrer Höhe mittel bis gering. Dies deutet darauf hin, dass unterschiedliche Aspekte der Rezeption von Testrückmeldungen erfasst wurden. Damit bietet sich für die Prüfung von Hypothese 3, dass Lehrkräfte in Thüringen unter Berücksichtigung von Kontextvariablen die Testrückmeldungen positiver bewerten als Lehrkräfte in Baden-Württemberg, eine multivariate Varianzanalyse an. Dieses Verfahren reduziert redundante Informationen aufgrund der Abhängigkeit der abhängigen Variablen. Es wird geprüft, ob die Unterschiede zwischen den Untersuchungsgruppen (Land, Fach, Schulform) bezüglich aller vier abhängigen Variablen signifikant sind.

Tabelle 6: Rezeption und Nutzung der Tests – Zusammenhänge zwischen den abhängigen Variablen

		Lehrplanvalidität	Hinweise Diagnostik	Hinweise Noten	Negative Folgen
Lehrplanvalidität	*r*	–	.48 ***	.35 ***	-.22 ***
	n		1749	1748	1746
Hinweise Diagnostik	*r*		–	.54 ***	-.28 ***
	n			1768	1752
Hinweise Noten	*r*			–	-.11 ***
	n				1751
Negative Folgen	*r*				–
	n				

Anmerkung. Zweiseitige Signifikanztests.
* $p < .05$, ** $p < .01$, *** $p < .001$.

Um ein möglichst sparsames Modell zu erhalten und die Stichprobengröße nicht zu reduzieren, werden in der multivariaten Varianzanalyse alle in der vorangehenden multiplen Regressionsanalyse nicht signifikanten Variablen ausgeschlossen (Alter, Beschäftigungsumfang und kollegiale Kooperation). Ebenso wird die Anzahl der Kovariaten durch die Bildung einer Indexvariable „Diskussion der Tests in Gremien“ noch einmal deutlich reduziert. Weitere Kovariaten sind dann noch

Schulgröße, Klassengröße und Schulleitung. Abhängige Variablen sind die vier Indikatoren für das Nutzungsverhalten (Curriculare Validität, Diagnosehinweise, Nutzung für Notengebung und negative Aspekte zentraler Tests). Als Faktoren wurden die Variablen Land, Schulform und getestetes Fach gewählt. Die einzelnen Zellen des Modells sind nicht gleich häufig besetzt, weil die baden-württembergische Stichprobe deutlich größer ist als die Zahl der in Thüringen befragten Lehrkräfte und weniger Lehrkräfte zu den Vergleichsarbeiten in den Fremdsprachen befragt wurden als zu den Vergleichsarbeiten in Deutsch oder Mathematik.

Tabelle 7: Effekte der Kontextvariablen auf die Rezeption und Nutzung von Tests – Ergebnisse des multivariaten Tests (n=1545)

	F (Pillai-Spur)	Hypothese df	Fehler df	Sig.	Partielles Eta-Quadrat
Konstanter Term	136.07	4	1525	.000	.263
Kovariaten					
Schulgröße	6.12	4	1525	.000	.016
Klassengröße	6.74	4	1525	.000	.017
Schulleitung (ja/nein)	4.12	4	1525	.001	.011
Lehrerselbstwirksamkeitserwartung	2.89	4	1525	.021	.008
Diskussion der Tests in Gremien	6.71	4	1525	.000	.017
Faktoren					
Land	23.06	4	1525	.000	.057
Schulform (Gymnasium)	4.17	4	1525	.002	.011
Fach	1.67	8	3052	.101	.004
Interaktionseffekte zwischen Faktoren					
Land x Schulform	6.21	4	1525	.000	.016
Land x Fach	3.33	8	3052	.001	.009
Schulform x Fach	1.13	8	3052	.339	.003
Land x Schulform x Fach	1.04	8	3052	.406	.003

Tabelle 7 zeigt die Ergebnisse des multivariaten Tests mit den Effekten der Kovariaten, Faktoren und Faktoreninteraktionen. Weil bei multivariaten Varianzanalysen Kovariaten und Faktoren unabhängig sein sollten, wurden diese Interaktionseffekte ebenfalls geprüft, jedoch nicht tabellarisch dargestellt. Die beiden Faktoren Land und Fach haben keine signifikante Interaktion mit den fünf Kovariaten. Allerdings verletzen zwei der fünf Interaktionsterme für den Faktor Schulform diese Voraussetzung. Es gibt signifikante Interaktionen zwischen Schulform und Schulgröße (F = 5.14; df = 4; part. Eta-Quadrat = .013; p = .000) bzw. Lehrerselbstwirksamkeitserwartung (F = 3.40; df = 4; part. Eta-Quadrat = .009; p = .009). Unter den Kovariaten führen die Diskussion der Tests in

Gremien, Mitgliedschaft Schulleitung sowie Klassen- und Schulgröße zu signifikanten Zusammenhängen und bestätigen damit die Befunde der Regressionsanalysen. Lediglich die Kovariate Lehrerselbstwirksamkeit wird nicht mehr signifikant. Der Zusammenhang war jedoch bereits bei den vorausgehenden Regressionsanalysen sehr schwach.

Die multivariaten Tests zeigen signifikante Haupteffekte für die beiden Faktoren Land (Testsystem) und Schulform sowie signifikante Interaktionseffekte zwischen Land und Schulform bzw. Land und getestetem Fach, die in den nachfolgenden Analysen für jede abhängige Variable gesondert betrachtet werden. Mit einer mehr als doppelt so großen Stichprobe als in der Studie von Maier (2008b) und unter Kontrolle der Intensität der Diskussion von Testrückmeldungen in Gremien kann Hypothese 3 mit dieser multivariaten Varianzanalyse bestätigt werden. Allerdings zeigen der signifikante Haupteffekt sowie die Interaktionseffekte, dass dieser Befund ebenfalls differenziert zu betrachten ist.

Gesondert für die vier abhängigen Variablen werden die signifikanten Haupt- und Interaktionseffekte berichtet und graphisch dargestellt (Test der Zwischensubjekteffekte). Abbildung 2 zeigt eine bessere Bewertung der Lehrplanvalidität der Deutsch- und Mathematik-Kompetenztests durch Thüringer Lehrkräfte (Haupteffekt Land: $F = 22.07$; Eta-Quadrat = .014; $p = .000$). Jedoch bewerten baden-württembergische Lehrkräfte die curriculare Validität der Vergleichsarbeiten in der ersten Fremdsprache (in der Regel Englisch) deutlich höher als in den beiden anderen Fächern (Haupteffekt Fach: $F = 3.35$; Eta-Quadrat = .006; $p = .012$ sowie Interaktionseffekt Land x Fach: $F = 9.01$; Eta-Quadrat = .012; $p = .000$). Dies deutet darauf hin, dass die Einschätzung der Lehrplanvalidität vor allem mit den fachspezifischen Testaufgaben und weniger mit dem Rückmeldesystem bzw. der Implementationsstrategie zusammenhängt. Hier sind genauere, fachdidaktische Vergleichsanalysen der Testaufgaben von Interesse.

Abbildung 2: Einschätzung der Lehrplanvalidität von Vergleichsarbeiten nach Fach, Bundesland und Schulart

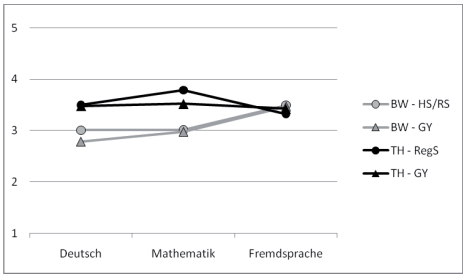
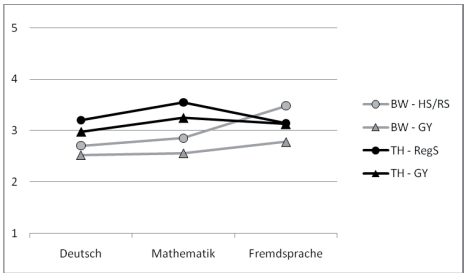


Abbildung 3: Einschätzung des diagnostischen Nutzens von Vergleichsarbeiten nach Fach, Bundesland und Schulart



Auch Abbildung 3 zeigt deutlich den signifikanten Ländereffekt bei der Einschätzung des diagnostischen Nutzens von Vergleichsarbeiten zugunsten der Thüringer Kompetenztests (F = 12.41; Eta-Quadrat = .008; p = .000). Für diese Variable gibt es ebenfalls einen weiteren Haupteffekt Fach (F = 4.42; Eta-Quadrat = .006; p = .012) und einen Interaktionseffekt Land x Fach (F = 5.76; Eta-Quadrat = .007; p = .003). Der dreifache Interaktionseffekt Land x Fach x Schulform wird allerdings nicht signifikant, auch wenn dies Abbildung 3 vermuten lässt (F = .96; Eta-Quadrat = .001; p = .385). D.h. die baden-württembergischen Lehrkräfte schätzen den diagnostischen Nutzen der Tests in der ersten Fremdsprache zum Teil höher ein als die Kollegen in Thüringen. Auch hier müssten weiterführende, fachdidaktische Analysen ansetzen, um diese Fächerdifferenz zu erklären. Dieser Befund weist jedoch darauf hin, dass man weniger von einem generellen Ländereffekt sprechen kann, sondern diesen auf die zentralen Tests in den beiden Fächern Deutsch und Mathematik eingrenzen sollte.

Abbildung 4: Einschätzung des Nutzens für die Notengebung von Vergleichsarbeiten nach Fach, Bundesland und Schulart

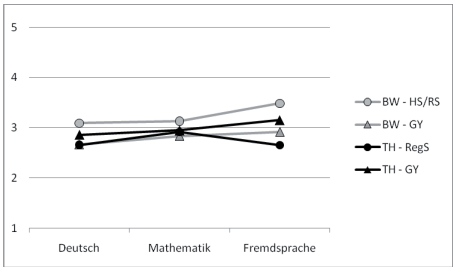
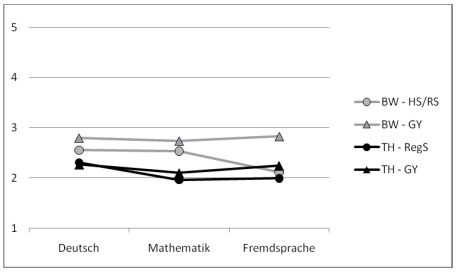


Abbildung 5: Einschätzung der negativen Folgen von Vergleichsarbeiten nach Fach, Bundesland und Schulart



Für die abhängige Variable „Einschätzung des Nutzens für die Notengebung“ gibt es sowohl einen signifikanten Haupteffekt Land (F = 17.73; Eta-Quadrat = .011; p = .000) als auch einen signifikanten Interaktionseffekt Land x Schulform (F = 16.83; Eta-Quadrat = .011; p = .000). Der Nutzen von Vergleichsarbeiten für die Notengebung wird vor allem von Haupt- und Realschullehrkräften in Baden-Württemberg höher eingeschätzt, weil hier während der ersten beiden Befragungswellen die Benotung der Tests möglich war und dies vermutlich einer Gruppe von Lehrkräfte eine interessante Vergleichsperspektive für ihre eigenen Bewertungsmaßstäbe eröffnete (Abbildung 4). Insgesamt sehen die Lehrkräfte kaum negative Folgen der zentralen Tests für ihren Unterricht oder die Zusammenarbeit im Kollegium (Abbildung 5). Die Lehrkräfte in Thüringen befürchten in signifikant geringerem Maße negative Folgen der zentralen Tests als ihre Kollegen in Baden-Württemberg (Haupteffekt Land: F = 21.27; Eta-Quadrat = .014; p = .000).

5. Diskussion

Mit dieser Studie sollte geprüft werden, ob die Länderdifferenzen zwischen Lehrkräften in Thüringen und Baden-Württemberg hinsichtlich der Bewertung von Vergleichsarbeiten stabil bleiben, wenn mit einer umfangreicheren Stichprobe ein weiteres Fach (erste Fremdsprache) und weitere Kontextfaktoren auf Schulebene kontrolliert werden (Hypothese 3). Die günstigeren Bewertungen der curricularen Validität und des diagnostischen Nutzens von Vergleichsarbeitsergebnissen durch Lehrkräfte in Thüringen wurden bestätigt und zwar sowohl durch die starken Effekte der unbereinigten Länderdifferenzen als auch durch die multivariate Varianzanalyse unter Kontrolle weiterer Kontextvariablen, vor allem der Diskussion von Testrückmeldungen in schulischen Gremien und der Tests in der ersten Fremdsprache. In Baden-Württemberg wurden dagegen die Nutzung der Testrückmeldungen für die Notengebung sowie die negativen Konsequenzen der Tests höher eingeschätzt. Damit bleiben die Befunde der Vorstudie (Maier, 2008b) durch die Vergrößerung der Stichprobe und die Berücksichtigung weiterer Untergruppen stabil.

Im Vergleich zur Vorstudie eröffnen diese Analysen eine weitere Erklärungsperspektive für die vorgefundenen Länderdifferenzen. Ein Problem dieses Ländervergleichs ist, dass die Differenzen in der Bewertung von zentralen Tests auch auf andere, nicht erfasste, länderspezifische Kontextvariablen zurückgeführt werden könnten. Beispielsweise liegt die Vermutung nahe, dass die allgemeine Zufriedenheit der Lehrkräfte mit der Bewertung bildungspolitischer Reformen interagiert. Dieser Effekt steckt vermutlich hinter den Zusammenhängen zwischen beispielsweise Klassengröße, Schulgröße, Schulfach und dem von den Lehrkräften eingeschätzten Rezeptionsverhalten. D.h. es gibt, wie im Rahmenmodell von Visscher und Coe (2003) beschrieben, Mediatorvariablen, die den Effekt des Testsystems auf die Bewertung der Lehrkräfte vermitteln. Unabhängig davon bleibt dennoch ein direkter Effekt des Testsystems auf die Bewertung der Testrückmeldungen durch die Lehrkräfte. Damit ist natürlich noch nicht eindeutig geklärt, auf welche Merkmale der Testsysteme oder der Implementation der Testsysteme diese unterschiedlichen Bewertungen zurückgehen. Die Variable „Land“ bleibt weiterhin eine „Omnibusvariable“, die auch sämtliche weiteren, länderspezifischen Differenzen zwischen Thüringen und Baden-Württemberg, wie z.B. Schulstruktur, bildungspolitisches Innovationsklima, Sozialisation der Lehrkräfte, etc. transportiert.

Die erste Hypothese, dass sich das Ausmaß der kollegialen Kooperation in Schulen positiv auf die Bewertung der Rezeption und Nutzung von Vergleichsarbeitsergebnissen durch einzelne Lehrkräfte auswirkt, muss eingeschränkt werden. Dies könnte damit zusammenhängen, dass sich die kollegiale Kooperation – wenn sie stattfindet – auf die Vorbereitung von Unterricht oder den Austausch von Lernmaterialien beschränkt. Hypothese 2, dass sich die Diskussion über Vergleichsarbeiten in schulischen Gremien auf die Rezeption und Nutzung der Daten

auf Lehrerebene auswirkt, muss spezifiziert werden. Die Daten geben Hinweise darauf, dass in Abhängigkeit des Gremiums spezifische Aspekte diskutiert werden und sich auf bestimmte Aspekte des Rezeptionsverhaltens von Lehrkräften auswirken können. Das Ausmaß der Diskussionen in Fachkonferenzen hängt beispielsweise mit der Bewertung des diagnostischen Nutzens zusammen, d.h. Lehrkräfte könnten durch ihre Fachkollegen stimuliert werden, sich genauer mit den Testergebnissen zu beschäftigen. In Klassenkonferenzen steht dagegen vermutlich eher der Nutzen der Tests für die Notengebung im Vordergrund.

Durch die Berücksichtigung von Vergleichsarbeiten in der ersten Fremdsprache ergab sich eine weitere, fach- und länderspezifische Differenz. Lehrkräfte in Baden-Württemberg bewerten die curriculare Validität und den diagnostischen Nutzen der Englisch-Vergleichsarbeiten deutlich besser als die Tests in den anderen Fächern. Es sollte jedoch einschränkend berücksichtigt werden, dass der relativ hohe Grad an Freiwilligkeit bei der Durchführung der Vergleichsarbeiten in der ersten Fremdsprache in Baden-Württemberg eventuell zu einer positiven Verzerrung der Lehrerbewertung beigetragen hat.

Betrachtet man die gesamte Varianzaufklärung, ergibt sich eine weitere Einschränkung der Befunde. Es ist anzunehmen, dass vor allem relevante Mediatorvariablen auf Lehrerebene auch in dieser Studie nicht hinreichend erfasst wurden. In weiteren Analysen wäre zu prüfen, inwiefern Einstellungen gegenüber technologischen Innovationen im Unterricht oder Überzeugungen zum pädagogischen Wert von Leistungsmessungen für einen großen Teil der hier nicht aufgeklärten Varianz verantwortlich sind. In der dritten und vierten Befragungswelle wurden entsprechende Skalen in den Fragebogen aufgenommen und es gibt deutliche Hinweise, dass das Ausmaß der Individualisierung des Unterrichts sehr hoch mit der Bewertung des diagnostischen Nutzens von Testrückmeldungen zusammenhängt.

Generell kann an dieser Form der Rezeptionsforschung kritisiert werden, dass die im theoretischen Teil dargestellten Annahmen zur Erklärung von Prozessen und Effekten datenbasierter Unterrichtsentwicklung nur ausschnitthaft empirisch modelliert wurden. Diese Problematik spiegelt sich bereits im Rahmenmodell von Visscher und Coe (2003), das die Befunde einer Vielzahl von mehr oder weniger vergleichbaren Studien zu dieser Thematik lediglich begrifflich ordnet. Es gibt bisher weder ein theoretisch stringentes Modell der Datennutzung auf Schul- und Unterrichtsebene, noch ein praktikables Forschungsdesign, um Effekte spezifischer Testrückmeldungen auf Unterrichtsqualität und Schülerleistungen nachzuweisen. Vor diesem Hintergrund sind auch die in dieser Studie geprüften Hypothesen lediglich der Versuch, einige der als relevant eingeschätzten Aspekte für die datenbasierte Unterrichtsentwicklung zu modellieren und empirisch zu prüfen. Hinzu kommt die Problematik, dass Merkmale des Testsystems nicht experimentell variiert werden können, sondern lediglich die konkret vorgefundenen Differenzen genutzt werden können, um mögliche Effekte von Testrückmeldungen abzuschätzen.

Diese Abschätzung lässt sich abschließend folgendermaßen zusammenfassen und kritisch einordnen: Die theoretisch begründbaren Differenzen zwi-

schen zwei Rückmeldesystemen innerhalb Deutschlands korrespondieren auch unter Kontrolle weiterer Mediatorvariablen mit statistisch signifikanten Differenzen in der Lehrerbewertung der Testrückmeldungen. Damit führt diese auf einer umfangreicheren Stichprobe basierende, erweiterte Analyse zu einer Bestätigung der Befunde der vorausgehenden Studie. In Thüringen scheint ein in weitere Schulentwicklungsstrategien eingebettetes Testsystem mit fairen Landesvergleichswerten auf relativ günstige Bedingungen an Schulen zu treffen und führt damit eher zu einem zielkonformen Rezeptionsverhalten der Testrückmeldungen als in Baden-Württemberg.

Die Varianzaufklärung durch die Länderdifferenz liegt allerdings im 1 %-Bereich, d.h. das Testsystem macht zwar einen Unterschied, aber es gibt gleichzeitig eine große Überlappung der Lehrerbewertungen in beiden Bundesländern. Dies ist vor allem der Tatsache geschuldet, dass die in dieser Studie gewählte Modellierung der Effekte im Vergleich zu den in der Literatur diskutierten Einflussvariablen unterkomplex ist. Ganze Variablenkomplexe zur Beschreibung eines zentralen Testsystems werden nicht erfasst (z.B. Aspekte der regionalen Implementierung von zentralen Tests) und es wird von linearen, nicht moderierten Wirkungen auf die schulinterne Nutzung ausgegangen. In komplexen Schul- und Unterrichtsentwicklungsprozessen spielen jedoch weitere dynamische Prozesse eine große Rolle. Weder die hier vorliegende Studie, noch die in der Literatur diskutierten Modelle liefern hierzu eine adäquate Modellierung.

Hinzu kommt, dass das Rahmenmodell (Visscher & Coe, 2003; Verhaeghe et al., 2010) nicht nur auf Differenzen zwischen den Testsystemen aufmerksam macht, sondern zahlreiche Aspekte auch weitreichende Gemeinsamkeiten der in Deutschland implementierten Testsysteme reflektieren. Weder in Baden-Württemberg, noch in Thüringen werden beispielsweise die Lehrkräfte bei der Auswahl der zu testenden Fachinhalte beteiligt. Die Lehrkräfte und Schulen können weder Testzeitpunkt noch die zur Verfügung gestellten Analysen selbst mitgestalten. In beiden Bundesländern werden zentrale Tests mit einem bürokratischen „Gestus“ angeordnet. Aufgrund von qualitativen Studien (z.B. Kuper & Hartung, 2007; Maier, 2009) ist jedoch zu vermuten, dass gerade diese Gemeinsamkeiten die Akzeptanz und Nutzung zentraler Testdaten schmälern können. Forschungsmethodisch bedeutet dies, dass vergleichende bzw. quasi-experimentelle Studien zwischen sehr unterschiedlich konzipierten Testsystemen einen wesentlich höheren Erklärungswert hätten. Dies wäre allerdings eher mit international vergleichenden Studien möglich. Als Kontrastfolie würden sich beispielsweise Länder wie Neuseeland[1] oder Schottland[2] anbieten. Dort werden momentan *assessment*-Systeme implementiert, die eine zentrale Festlegung und Prüfung von Bildungsstandards mit Lehrerfortbildungen zur formativen Leistungsdiagnostik sowie zur Praxis der Notengebung kombinieren.

1 http://nzcurriculum.tki.org.nz/National-Standards.
2 http://www.ltscotland.org.uk/.

Literatur

Baumert, J. (2001). Vergleichende Leistungsmessung im Bildungsbereich. *Zeitschrift für Pädagogik, 43. Beiheft*, 13–36.

Blum, W., Drüke-Noe, C., Leiß, D., Wiegand, B. & Jordan, A. (2005). Zur Rolle von Bildungsstandards für die Qualitätsentwicklung im Mathematikunterricht. *Zentralblatt für Didaktik der Mathematik, 37*, 267–274.

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

Büchter, A. & Leuders, T. (2005a). Quality development in mathematics education by focussing on the outcome: New answers or new questions? *Zentralblatt für Didaktik der Mathematik, 37*, 263–266.

Büchter, A. & Leuders, T. (2005b). From students' achievement to the development of teaching: Requirements for feedback in comparative tests. *Zentralblatt für Didaktik der Mathematik, 37*, 324–334.

Cheng, L. (1999). Changing assessment: Washback on teacher perspectives and actions. *Teaching and Teacher Education, 15*, 253–271.

Cheng, L. & Curtis, A. (2004). Washback or backwash: A review of the impact of testing on teaching and learning. In L. Cheng, Y. Watanabe & A. Curtis (Hrsg.), *Washback in language testing. Research contexts and methods* (S. 3–17). London: Lawrence Erlbaum.

Demie, F. (2003). Using value-added data for school self-evaluation: A case study of practice in inner-city schools. *School Leadership & Management, 23,* 445–467.

Ditton, H., Arnoldt, B. & Bornemann, E. (2002). Entwicklung und Implementation eines extern unterstützten Systems der Qualitätssicherung an Schulen – QuaSSU. *Zeitschrift für Pädagogik, 45. Beiheift*, 374–389.

Ditton, H. & Merz, D. (2000). *Qualität von Schule und Unterricht – Kurzbericht über erste Ergebnisse einer Untersuchung an bayerischen Schulen*. Katholische Universität Eichstätt/Universität Osnabrück. Zugriff am 03.01.2003 unter http://www.quassu.net/index.htm

Firestone, W. A., Winter, J. & Fitz, J. (2000). Different assessments, common practice? Mathematics testing and teaching in the USA and England and Wales. *Assessment in Education, 7,* 13–37.

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

Harris, A., Jamieson, I. & Russ, J. (1995). A study of 'effective' departments in secondary schools. *School Organisation*, *15* (3), 283–299.

Hayes, S. G. & Rutt, S. (1999). Primary analysis for secondary schools: An LEA research officer's perspective on helping secondary schools interpret assessment data for school improvement purposes. *Improving Schools, 2*, 44–52.

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

Herzog, W. (2010). Besserer Unterricht dank Bildungsstandards und Kompetenzmodellen? In A. Gehrmann, U. Hericks & M. Lüders (Hrsg.), *Bildungsstandards und Kompetenzmodelle: Beiträge zu einer aktuellen Diskussion über Schule, Lehrerbildung und Unterricht* (S. 37–46). Bad Heilbrunn: Klinkhardt.

Hosenfeld, I. & Groß Ophoff, J. (2007). Nutzung und Nutzen von Evaluationsstudien in Schule und Unterricht. *Empirische Pädagogik, 21*, 352–457.

Klieme, E. (2004). Begründung, Implementation und Wirkung von Bildungsstandards: Aktuelle Diskussionslinien und empirische Befunde. *Zeitschrift für Pädagogik, 50*, 625–634.

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

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

Louis, K. S., Febey, K. & Schroeder, R. (2005). State-mandated accountability in high schools: Teachers' interpretations of a new era. *Educational Evaluation and Policy Analysis, 27*, 177–204.

Maier, U. (2008a). Rezeption und Nutzung von Vergleichsarbeiten – Ergebnisse einer Lehrerbefragung in Baden-Württemberg. *Zeitschrift für Pädagogik, 54*, 95–117.

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*. Hohengehren: Schneider.

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.

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*. Zürich: Bildungsdirektion des Kantons Zürich.

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

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

Opfer, V. D., Henry, G. T. & Mashburn, A. J. (2008). The district effect: Systemic responses to high stakes accountability policies in six southern states. *American Journal of Education, 114*, 299–332.

Pant, H. A., Vock, M., Pöhlmann, C. & Köller, O. (2008). Offenheit für Innovation. Befunde aus einer Studie zur Rezeption der Bildungsstandards bei Lehrkräften und Zusammenhänge mit Schülerleistungen. *Zeitschrift für Pädagogik, 54*, 827–845.

Peek, R. & Dobbelstein, P. (2006). Benchmarks als Input für die Schulentwicklung – das Beispiel der Lerstandserhebungen in Nordrhein-Westfalen. In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen* (S. 41–58). Münster: Waxmann.

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

Peng, W. J., Thomas, S. M., Yang, X. & Li, J. (2006). Developing school evaluation methods to improve the quality of schooling in China: A pilot 'value added' study. *Assessment in Education, 13*, 135–154.

Rossi, P. H., Lipsey, M. & Freeman, H. E. (2004). *Evaluation: A systematic approach*. London: Thousand Oaks.

Rudd, P. & Davies, D. (2002). *A revolution in the use of data? The LEA role in data collection, analysis and use and its impact on pupil performance*. Slough: NFER.

Saunders, L. (2000). Understanding schools' use of 'value added' data: The psychology and sociology of numbers. *Research Papers in Education, 15*, 241–258.

Schmitz, G. S. & Schwarzer, R. (2000). Selbstwirksamkeitserwartungen von Lehrern: Längsschnittbefunde mit einem neuen Instrument. *Pädagogische Psychologie, 14,* 12–25.

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: Freie Universität.

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

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

Verhaeghe, G., Vanhoof, J., Valcke, M. & Van Petegem, P. (2010). Using school performance feedback: Perceptions of primary school principals. *School Effectiveness and School Improvement, 21*, 167–188.

Visscher, A. J. & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School effectiveness and school improvement, 14*, 321–349.

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

Wikeley, F., Stoll, L. & Lodge, C. (2002). Effective school improvement: English case studies. *Educational Research and Evaluation. School effectiveness and school improvement in a European context, 8*, 363–385.

Yang, M., Goldstein, H., Rath, T. & Hill, N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education, 25*, 469–483.

Anhang

Tabelle 8: Skalen zur Erfassung der Rezeption und Nutzung zentraler Tests – Hauptkomponentenanalyse mit Varimax-Rotation; 4 Faktoren mit Eigenwert > 1 erklären 66,5 % der Gesamtvarianz

	Komponente			
	1	2	3	4
Tests helfen Stärken und Schwächen der Schüler zu benennen (dia9)	.78			
Tests geben zusätzliche diagnostische Hinweise (dia10)	.76			
Testrückmeldung ist Grundlage für die Planung individueller Fördermaßnahmen (dia5)	.75			
Tests sind für die Arbeit der Schule wichtig (su4)	.74			
Testrückmeldung ist tragfähige Argumentationsbasis für Beratungsgespräche mit Eltern (dia7)	.73			
Tests sollen weiterhin regelmäßig durchgeführt werden (su3)	.71		-.31	
Testaufgaben decken die Lernbereiche des Bildungsplanes ab (auf1)		.83		
Testaufgaben stimmen mit der Gewichtung der Lernbereiche im Lehrplan überein (auf2)		.81		
Testaufgaben entsprechen dem im Lehrplan geforderten Leistungsniveau (auf5)		.75		
Test übt zusätzlichen Druck auf Schulen und Lehrer aus (su10)			.81	
Test führt dazu, dass nur noch für den Test geübt wird (su11)			.79	
Test führt zu Konkurrenz und Missgunst zwischen den Lehrern (su9)			.77	
Tests regen zum Nachdenken über eigene Bewertungsmaßstäbe an (dia17)				.83
Testrückmeldung gibt Hinweise, ob eigene Klassenarbeiten zu schwer oder leicht sind (dia6)				.79
Testrückmeldung in die Jahresendnote einfließen zu lassen ist sinnvoll (dia16)				.63

Anmerkung. Faktorenladungen unter .30 werden nicht dargestellt.