Welche Konsequenzen ziehen Mathematiklehrkräfte aus verpflichtenden Diagnose- und Vergleichsarbeiten?

Ergebnisse einer Interviewstudie

von

Uwe Maier, Schwäbisch Gmünd

Kurzfassung: Landesweite, standardbasierte Leistungsmessungen sollen eine datenbasierte Weiterentwicklung von Unterricht anregen und unterstützen. In einer Interviewstudie wurde deshalb untersucht, welche Konsequenzen Mathematiklehrkräfte an Grund- und Sekundarschulen (n = 50) aus den ersten baden-württembergischen Diagnose- und Vergleichsarbeiten für ihren zukünftigen Unterricht gezogen haben. Die qualitativ-inhalts-analytische Auswertung der Lehreraussagen zeigte, dass zentrale Tests am Ende eines zweijährigen Bildungsabschnitts ein zu oberflächliches Maß sind, um den mathematischen Kompetenzaufbau in verschiedenen Inhaltsbereichen abzubilden. Nur eine Minderheit der befragten Lehrkräfte wurde durch einzelne Aufgabenstellungen dazu angeregt, über den vorausgehenden Unterricht nachzudenken und mögliche Veränderungen abzuleiten. Abschließend wird diskutiert, inwiefern die spezifische Anlage der baden-württembergischen Diagnose- und Vergleichsarbeiten für das Befragungsergebnis verantwortlich gemacht werden kann.

Abstract: This paper analyzes how and to what extent math teachers draw instructional conclusions from state mandated school performance tests that aim to encourage data-based school improvement. The interview study approached elementary and secondary school teachers (n = 50) in Baden-Württemberg, one of the larger German states. Mostly sobering results of the qualitative content analysis show that a single math test after a two-year learning period cannot provide sufficient information for better understanding students' mathematical knowledge acquisition. Only a minority of the teachers explained that some test assignments prompted reflection about slightly changing their instructional approach. The paper concludes with a discussion on the characteristics of the testing and feedback system in Baden-Württemberg and its potential impact on the schools' disapproval to make sensible use of the test results.

Bildungspolitiker und Protagonisten einer standardbasierten Schulreform versprechen sich von zentralen Leistungsmessungen eine verbesserte schulinterne Qualitätssicherung. Vergleichsarbeiten sollen unter anderem zu einer stärkeren Orientierung des Unterrichts an staatlichen Standards beitragen (KMK 2003; Klieme et al.

2003). Empirische Studien in Ländern mit langjähriger Testtradition und erste Befunde aus dem deutschsprachigen Raum geben allerdings Anlass, diese Reformeuphorie wesentlich kritischer zu sehen. Es stellt sich die Frage, wie Testrückmeldungen von Lehrern[1] interpretiert werden, und ob dies tatsächlich zu besseren Lernergebnissen führt. Dieser Artikel berichtet Ergebnisse einer Interviewstudie mit Lehrkräften, die direkt im Anschluss an die ersten baden-württembergischen Vergleichsarbeiten 2006 durchgeführt wurde. In der Befragung ging es unter anderem darum, welche Konsequenzen Lehrkräfte aus den Rückmeldungen für den eigenen Mathematikunterricht ziehen.

Im einleitenden Teil des Artikels wird zunächst einmal ein Überblick zur Forschungslage gegeben. Vor allem in den traditionell testdominierten Bildungssystemen (USA, England) wurden die Auswirkungen zentraler Leistungsmessungen auf Unterricht breit erforscht. Aber auch im deutschsprachigen Raum gibt es mittlerweile erste Ergebnisse zur Nutzung landesweiter Tests durch Lehrkräfte. Anschließend werden methodische Vorgehensweise und Datenauswertung der Interviewstudie beschrieben. Der Ergebnisteil bündelt zentrale und häufig vorkommende Aussagen der befragten Lehrkräfte. Lehrerzitate veranschaulichen die gefundenen Aussagekategorien. In einem abschließenden Teil des Artikels werden mögliche Konsequenzen für die Weiterentwicklung der baden-württembergischen Vergleichsarbeiten diskutiert.

1 Problemstellung

Zentrale Leistungsmessungen als bildungspolitisches Multifunktionswerkzeug haben in Ländern wie den USA oder Großbritannien bereits eine jahrzehntelange Tradition. Aus administrativer Perspektive möchte man das Schulsystem mit folgender Handlungslogik steuern: Setze hohe Standards, teste die Lernergebnisse der Schüler gegen diese Standards und gib Rückmeldungen an Schulen und Lehrkräfte, wie groß die Lücke zwischen Zielen und Leistungen ist. Verknüpfe die Rückmeldung mit Belohnungen und Sanktionen. Stillschweigend wird angenommen, dass sich Schulen und Lehrer auf diese Weise zu verstärkten Anstrengungen motivieren lassen und die Leistungsrückmeldungen tatsächlich auch zu einer datengestützten Verbesserung des Unterrichts nutzen.

Die Schwachstellen dieser bildungspolitischen Handlungslogik wurden in zahlreichen empirischen Studien aufgedeckt (z. B. Amrein & Berliner 2002). Besonders negative Auswirkungen auf Unterricht haben sog. „high-stakes tests“ im Rahmen

[1] Aus Gründen der Lesbarkeit werden die männlichen Formen „Lehrer“ und „Schüler“ verwendet. Gemeint sind damit beide Geschlechter.

von Testkonzeptionen, die in der Literatur als „minimum competency testing“ zusammengefasst wurden (Stecher 2002; Herman 2004; Amrein & Berliner 2003). Die Idee dieser Testprogramme war, grundlegende Basiskompetenzen in den zentralen Fächern abzuprüfen. Schulen und Lehrkräfte sollten durch entsprechende Sanktionen dafür verantwortlich gemacht werden, dass alle Kinder die notwendigen Grundqualifikationen erreichen. Ohne auf den breiteren pädagogischen Diskurs einzugehen, werden die wichtigsten Auswirkungen auf Unterricht stichwortartig skizziert:

- Schulen, die aufgrund schlechter Testwerte von Sanktionen bedroht sind, reduzieren ihr Curriculum auf testrelevante Fächer.
- Auch innerhalb der getesteten Fächer werden die Lerninhalte deutlich eingeschränkt; z. B. Reduktion des Mathematikunterrichts auf Basisfertigkeiten.
- Durch unterschiedliche Strategien bzw. „Schummeleien“ werden die Testergebnisse manipuliert.
- Schwache Schüler werden vor dem Test vom Unterricht ausgeschlossen bzw. nur Schüler mit „Steigerungspotenzial“ erhalten eine Lernförderung.
- Sinkende Motivation bei Schülern und Zunahme der Schulabbrecherquoten.

Die massive fachdidaktische und erziehungswissenschaftliche Kritik war in einigen US-Bundesstaaten Anlass für grundlegende Änderungen mit dem Ziel, anspruchsvollere Kompetenzen in den jeweiligen Fächern zu messen. Beispielsweise wurde der Anteil simpler Multiple-Choice-Items zur Überprüfung von Faktenwissen reduziert. Dafür wurde mit sog. „performance tests“ oder „portfolio assessments“ experimentiert, um eher komplexe, mathematische Kompetenzen prüfen zu können (McDonnell & Choisser 1997). Popham (1987) fasst diese neue Handlungslogik unter dem Begriff „measurement-driven instruction“ zusammen. Und auch diesmal versprach sich die Bildungspolitik eine positive Beeinflussung der Unterrichtspraxis.

In empirischen Studien konnten ebenfalls Auswirkungen der neuen Testsysteme auf Unterricht nachgewiesen werden. In Vermont und Kentucky orientierten sich Lehrkräfte verstärkt an den im Test verwendeten Problemlöseaufgaben und ließen sich zur Nutzung vielfältiger mathematischer Repräsentationen anregen (Koretz et al. 1994). In Vermont wurde ein Portfolio-Testprogramm eingeführt, das die Lehrer dazu brachte, eine bestimmte Art von mathematischen Problemlöseaufgaben im Unterricht verstärkt zu behandeln. Dies führte zwar insgesamt zu besseren Testleistungen, hatte allerdings aufgrund der Spezifität der Problemstellungen wenig mit der Entwicklung einer allgemeinen mathematischen Problemlösefähigkeit zu tun (Stecher & Mitchell 1995).

Die Effekte waren zunächst vielversprechend, allerdings auch in ihrer Reichweite begrenzt. Firestone, Winter und Fitz (2000) verglichen die Auswirkungen einer neuen Generation von „performance-based tests“ in den US-Bundesstaaten Maine und Maryland sowie in England und Wales. Die neuen Tests beeinflussen vor allem die Übungsphasen während der Testvorbereitung, die Unterrichtsmethoden scheinen sich nicht zu ändern. Lehrer greifen immer wieder auf komplexere Problemlöseaufgaben zurück, versäumen es jedoch deren Potenzial voll auszuschöpfen. Beispielsweise wurde die Anzahl möglicher Lösungen durch Vorgaben schon im Vorfeld eingeschränkt. Die Änderungen werden insgesamt als oberflächlich bezeichnet und beziehen sich wiederum gezielt auf die Testvorbereitung. Eine prinzipielle Änderung der Vorgehensweise im Mathematikunterricht auf breiter Basis konnte in keinem der untersuchten Länder festgestellt werden.

Die deutsche Bildungspolitik kopiert zumindest ansatzweise diese testbasierten Reformbemühungen. Die Zielsetzungen und rhetorischen Muster gleichen sich. Es gibt allerdings einen fundamentalen Unterschied zu anglo-amerikanischen „accountability systems“: Die Testergebnisse werden noch nicht öffentlich publiziert und die Konsequenzen für leistungsschwache Schulen sind sehr gering und indirekt. Der Handlungsdruck für Lehrkräfte und Schulen ist momentan somit wesentlich geringer. In einigen Bundesländern gibt es jedoch schon konkrete Pläne, die Testergebnisse schrittweise auch auf Schulebene zu veröffentlichen (z. B. Nordrhein-Westfalen, Mecklenburg-Vorpommern). Dies würde dann zu einem erhöhten indirekten Handlungsdruck auf Lehrkräfte führen. Die Konsequenzen lassen sich aufgrund der internationalen Befunde leicht antizipieren.

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; Peek & Nilshorn 2004). Dabei zeigte sich durchweg, dass die Rückmeldung hoch aggregierter Daten aus Stichprobenstudien von Lehrkräften nicht mit dem eigenen Unterricht in Verbindung gebracht wurde und ja auch nie primäres Ziel dieser Studien war. Wenn überhaupt eine Rezeption der Ergebnisse stattfand, dann noch am ehesten in den Mathematik-Fachkonferenzen. Eine schulweite Diskussion konnte nicht beobachtet werden.

Die nach und nach von den einzelnen Bundesländern eingeführten Vergleichsarbeiten für die Hauptfächer Deutsch und Mathematik haben – im Vergleich zu „large scale assessments“ – ganz klar die Zielsetzung, den schulinternen Diskurs über Leistungsergebnisse und mögliche Ursachen anzuregen. Aus diesem Grund wurden im Rahmen dieser Vergleichsarbeitsprojekte die Rezeptionsstudien fortgeführt und methodisch weiterentwickelt (Nachtigall 2005; Groß Ophoff et al. 2006; Bonsen, Büchter & Peek 2006; Sill & Sikora 2007). Auch aus der Schweiz sind bereits

entsprechende Untersuchungsergebnisse bekannt (Moser 2003; Keller & Moser 2006; Tresch 2007; Baeriswyl et al. 2006). An dieser Stelle sollen erste Ergebnisse und bereits erkennbare Tendenzen der Rezeption von Leistungsrückmeldungen durch Mathematiklehrkräfte skizziert werden.

Beispielsweise wurden im Rahmen des Verbundprojektes VERA Grundschullehrkräfte online befragt, welche Konsequenzen sie aus den Leistungsrückmeldungen ziehen (Groß Ophoff et al. 2006). 70% der befragten Lehrkräfte gaben an, aufgrund der Testrückmeldung Inhalte zu wiederholen und zu vertiefen. Über die Hälfte der Lehrkräfte dachte kritisch über die eigenen Unterrichtsmethoden nach. Die Autoren sehen VERA ebenfalls als Ursache für eine intensivere schulinterne Kooperation. Nachtigall (2005) führte eine repräsentative Fragebogenstudie zur Evaluation der Nutzung von Leistungsrückmeldungen der Thüringer Kompetenztests 2004 durch. Dabei zeigte sich eine Erhöhung des Spektrums und der Anzahl der durch die Rückmeldungen eingeleiteten Folgemaßnahmen im Vergleich zur Nutzung der Kompetenztestergebnisse 2003.

Moser (2003) befragte Lehrerinnen und Lehrer im Kanton Zürich über erste Erfahrungen mit dem „Klassencockpit“. Das Klassencockpit wird von einem Lehrmittelverlag angeboten und ist ein freiwilliges Evaluationsinstrument, das standardisierte Leistungstests für verschiedene Unterrichtsinhalte in den Fächern Deutsch und Mathematik anbietet. Aufgrund der Freiwilligkeit des Angebots konnte Moser nach den Motiven für den Einsatz der Leistungstests fragen. Während fast alle Lehrkräfte an einem sozialen Vergleich mit den kantonalen Mittelwerten interessiert waren und viele die Ergebnisse auch als Bestätigung der eigenen Schülerbeurteilungen sehen, gibt nur jede dritte Lehrkraft an, aufgrund der Leistungsrückmeldung über den eigenen Unterricht reflektiert zu haben.

Keller und Moser (2006; vgl. auch Tresch 2007) führten eine schriftliche Lehrerbefragung zur Akzeptanz und Nutzung der im Schweizer Kanton Aargau eingeführten Leistungstests „Check 5“ durch und kommen bezüglich der Nutzungsoptionen zu gegensätzlichen Ergebnissen. Die überwiegende Mehrheit der befragten Lehrer gaben an, Maßnahmen aufgrund der Testrückmeldung ergriffen zu haben. Dagegen erwähnte nur ein Drittel der Befragten eine Anpassung des eigenen Beurteilungsmaßstabes auf Grundlage der Testergebnisse. Fast alle anvisierten Maßnahmen wurden von den Lehrkräften zum Zeitpunkt der Umfrage bereits voll bzw. teilweise umgesetzt.

In Mecklenburg-Vorpommern untersuchten Sill und Sikora (2007) die neu eingeführten zentralen Mathematiktests sowie den Umgang mit den Ergebnissen in den Pilotschulen. Schulleiter und Fachleiter in Mathematik wurden in Form eines freiwilligen Gesprächs über die Ergebnisse informiert. Die Auswertungsveranstaltungen und Gespräche an den Stichprobenschulen wurden qualitativ dokumentiert.

Dabei zeigte sich, dass bereits einfache statistische Darstellungen Verständnisschwierigkeiten bereiten können. Von großer Bedeutung war auch die Vertrauensbildung. Die Schulleiter waren nur dann zu einer offenen Diskussion bereit, wenn klar gemacht wurde, dass die Veranstaltungsergebnisse nicht an die Schulaufsicht weitergeleitet wurden. Mittelmäßige oder gute Ergebnisse waren für die Schulleiter zufriedenstellend und konnten keine weitergehende Reflexion anregen. Sehr schlechte Ergebnisse wurden überwiegend external attribuiert (Unterrichtsausfall, Schüler kannten Aufgabenformate nicht, etc.).

Eine weitere Schwierigkeit war, dass viele Lehrkräfte die Notwendigkeit einer vertiefenden Fehleranalyse nicht einsahen. Es war nicht unbedingt klar, dass man aus Fehleranalysen Schlussfolgerungen für den eigenen Unterricht ziehen kann. Lehrkräfte werteten die Vergleichsarbeit dagegen eher als summative Evaluation und behandelten sie wie Abschlussprüfungen. Bezogen auf die Zielsetzung der Unterrichtsentwicklung kommen Sill und Sikora (2007, 247) zum Schluss, „dass Lehrer gar nicht erwarten, dass aus Leistungsvergleichen Schlussfolgerungen auf den Prozess des Unterrichts gezogen werden können. Sie haben keine Methoden gelernt oder erprobt, die ihnen derartige Schlüsse erlauben." Eine gewisse Steuerungswirkung ergibt sich lediglich über die Aufgabenstellungen. Wie bei zentralen Abschlussprüfungen werden bestimmte Aufgabenformate in den Unterricht und in Klassenarbeiten übernommen und geben somit indirekt den Leistungsstandard vor.

Die bisher in Rezeptionsstudien herausgearbeiteten Defizite lassen sich gut mit Forderungen aus der Evaluationsforschung in Einklang bringen. Evaluationsergebnisse führen nur dann zu datenbezogenen Verbesserungen, wenn alle Betroffenen bei der Planung, Durchführung und Ergebnisinterpretation beteiligt werden (Wottawa & Thierau 1998; Lind 2003). Bei der Vorgabe zentraler Vergleichsarbeiten wird genau dieses Kriterium verletzt, zumindest was die Planung und Durchführung anbelangt. Somit wird die Entwicklung einer gemeinsamen Verantwortlichkeit für das Lernen der Schüler in der Einzelschule sehr erschwert. Es kommt möglicherweise zu Solidarisierungseffekten zwischen Lehrern und Schülern und somit zu verfälschten Ergebnissen. Beispiele aus der Schweiz zeigen, dass freiwillige Teilnahme und Auswahlmöglichkeiten zu höherer Akzeptanz führen können.

2 Zentrale Tests in Baden-Württemberg und Fragestellung der Studie

In Baden-Württemberg wurden gegen Ende des Schuljahres 2005/06 die ersten verpflichtenden Diagnose- und Vergleichsarbeiten vom Landesinstitut für Schulentwicklung in Stuttgart durchgeführt. Grundschüler wurden mit Diagnosearbeiten in den Fächern Deutsch und Mathematik am Ende der zweiten Klasse getestet. In

den weiterführenden Schulen heißen die zentralen Tests „Vergleichsarbeiten“ und wurden in den Fächern Deutsch und Mathematik am Ende der 6. Klasse geschrieben (in der Realschule und im Gymnasium auch in der 8. Klasse). Je nach Schulform standen Tests für weitere Fächer zur Verfügung. Offizielle Zielsetzungen der Tests sind Qualitätssicherung und objektive Lernstandsrückmeldung: Die Vergleichsarbeiten sollen „Lehrerinnen und Lehrern, den Schülerinnen und Schülern und deren Eltern objektive Informationen über den jeweiligen individuellen Lernstand im Hinblick auf bestimmte Kompetenzen“ geben und zur „zielgerichteten und systematischen Qualitätsentwicklung vor Ort“, d. h. zur Weiterentwicklung von Unterricht beitragen.[2]

Die Testaufgaben wurden von Kommissionen unter Leitung des Landesinstituts für Schulentwicklung in Stuttgart entwickelt, die unter anderem mit erfahrenen Lehrkräften sowie Psychometrikern besetzt waren. Die Diagnose- und Vergleichsarbeiten wurden ein Jahr vorher in einer Pilotstudie erprobt. Über die Auswahl der an der Pilotstudie beteiligten Schulen stehen keine öffentlich zugänglichen Informationen zur Verfügung. Die Pilotierungsergebnisse wurden als landesweite Vergleichswerte in selbstauswertenden Excel-Tabellen zur Verfügung gestellt.

Die Durchführung und Auswertung der Vergleichsarbeiten oblag komplett der einzelnen Schule. Über den Landesbildungsserver erhielten die Schulleitungen die Tests, die selbstauswertenden Excel-Tabellen sowie die Handreichungen. Direkt nach der Durchführung wurden die Testpunkte pro Aufgabe und Schüler direkt in die Excel-Tabellen eingegeben. Berechnet wurden die Punktwerte pro Schüler und Klasse sowie die Aufgabenschwierigkeiten pro Klasse. Die Aufgabenschwierigkeiten sowie das Klassenergebnis konnten sodann mit den Ergebnissen der Pilotierungsstudie verglichen werden. Konfidenzintervalle wurden allerdings nicht angegeben. Die Schüler- und Klassenpunkte wurden ebenfalls in „Leistungsstufen“ umgerechnet. Diese orientierten sich allerdings lediglich an der prozentualen Häufigkeitsverteilung der Pilotierungsergebnisse. Die Lehrkräfte waren verpflichtet, den Eltern Leistungsstufe und Punktwert des Schülers mitzuteilen.

Die Handreichungen für die Lehrkräfte waren folgendermaßen aufgebaut: Zunächst einmal wurden die Aufgabenlösungen und die Bezüge zum Bildungsplan aufgelistet. Danach wurde die Dateneingabe erläutert. Es folgte ein Abschnitt zur Erklärung der einzelnen Ergebnisgrafiken in den Excel-Tabellen. Für die Vergleichsarbeiten in der Sekundarstufe wird die Umrechnungstabelle von Punkten in Noten abgebildet. Zur Dateninterpretation wurden zahlreiche Anregungen auf einem sehr allgemeinen Niveau gegeben. Weitere fachdidaktische Kommentare zu

---

[2] Landesinstitut für Schulentwicklung (2006): Diagnosearbeiten 2006, Grundschule Klasse 2, Mathematik, Hinweise für die Lehrerin und den Lehrer.

den Aufgaben sowie spezifische Fehleranalysen fehlten. In Baden-Württemberg hat man sich aufgrund der (im Jahr 2005) noch unsicheren Befundlage im Bereich der Kompetenzmodellierung dafür entschieden, den Vergleichsarbeiten keine Kompetenzmodelle zu Grunde zu legen. Auch auf die fachdidaktische Kommentierung der Aufgaben wurde weitestgehend verzichtet.

Die Diagnosearbeit Mathematik in Klasse 2 deckte mit ihren Aufgaben sämtliche Leitideen im Bildungsplan Mathematik der Jahrgangsstufe 1/2 ab: Zahl, Messen und Größen, Raum und Ebene, Muster und Strukturen, Daten und Sachsituationen. In der Testbeschreibung wurden die Aufgabenstellungen jeweils mathematischen Kompetenzbereichen zugeordnet. Die Mehrzahl der Items prüfte grundlegende Basiskompetenzen im Bereich Rechnen ab (Leitidee Zahl). Mit einigen wenigen Aufgaben zu Sachsituationen wurden aber auch einfache mathematische Modellierungen gefordert. Die Mathematikvergleichsarbeiten in der Sekundarstufe waren analog aufgebaut und repräsentierten mit jeweils zwei bis drei Aufgaben im Bildungsplan vertretene Leitideen (z. B. in der Hauptschule: Zahl, Messen, Raum und Form, Funktionaler Zusammenhang, Modellieren, Daten und Zufall). Die Items prüften grundlegende arithmetische Kompetenzen, die Entnahme und Interpretation von Informationen aus Diagrammen und die mathematische Modellierungsfähigkeit von Sachsituationen.

Die Diagnose- und Vergleichsarbeiten können somit als eine summative Leistungsmessung am Ende eines im Bildungsplan ausgewiesenen Bildungsabschnitts bezeichnet werden. Was die Innovationsfähigkeit der Aufgaben anbelangt, so kann zumindest ansatzweise von neuen Impulsen geredet werden. Alle drei untersuchten Vergleichsarbeiten in Mathematik enthalten Aufgaben, bei denen Informationsentnahme und einfache mathematische Modellierungen gefordert werden. Der überwiegende Teil der Aufgaben prüft eher einfache Rechenalgorithmen ab.

Eine grundlegende Schwierigkeit lag allerdings im Durchführungszeitpunkt kurz vor den Sommerferien. Da Klassen in der Regel nach der zweiten bzw. 6. Klasse von den Fachlehrkräften abgegeben werden, konnte sich der Rückmeldeeffekt nur sehr begrenzt auf die Lerngruppe beziehen, die den Test geschrieben hatte. Die Testrückmeldungen konnten allenfalls im Zuge eines Übergabegesprächs an den nachfolgenden Kollegen weitergeleitet werden. Das bedeutet, wenn von den Mathematiklehrkräften Konsequenzen gezogen wurden, können sich diese lediglich auf den zukünftigen Unterricht mit einer anderen Klasse beziehen.

Unter diesen Ausgangsbedingungen stellt sich somit die Forschungsfrage, welche Konsequenzen aus den ersten Diagnose- und Vergleichsarbeiten in Baden-Württemberg für den zukünftigen Mathematikunterricht gezogen werden. Dabei ist von entscheidender Bedeutung, ob diese Konsequenzen auch tatsächlich Ergebnis einer Reflexion der Testergebnisse vor dem Hintergrund des eigenen Unterrichts sind.

Beispielsweise beschreiben Sill und Sikora (2007) die ursächliche Verknüpfung von Rückmeldung und Kompetenzaufbau im eigenen, vorausgegangenen Unterricht als eine zentrale Voraussetzung für die weitere, schulinterne Nutzbarmachung von Vergleichsarbeiten. Unerwünschte Nebenwirkungen zentraler Tests entstehen in der Regel dann, wenn sich die Reflexion über Testergebnisse vorrangig nicht auf den Kompetenzaufbau bezieht, sondern andere Bezugspunkte wie z. B. die Reputation als Lehrer im Vordergrund stehen.

Ebenso interessieren mögliche Bedingungen für die Nutzbarmachung von Testergebnissen. Können bestimmte Merkmale des Tests und einzelner Aufgabenstellungen beschrieben werden, die in besonderem Maße zu einer kritischen Reflexion anregen können? Oder gibt es bestimmte Aspekte des Testsystems, die eine Ableitung von Konsequenzen verhindern? Beispielsweise liegt die Vermutung nahe, dass die spärliche fachdidaktische Kommentierung und der Verzicht auf Kompetenzmodelle dazu führen, dass Lehrkräfte kaum einen Unterschied zu eigenen Klassenarbeiten erkennen können. Die Nutzbarmachung hängt aber auch von den Lehrkräften selbst ab. Hier ist zu fragen, welche Voraussetzungen Lehrkräfte für eine produktive Nutzung von Leistungsrückmeldungen im Mathematikunterricht mitbringen sollten.

3 Methodisches Vorgehen

Die hier vorgestellten Ergebnisse sind Teil einer Längsschnittuntersuchung, in der die Rezeption und Nutzung zentraler Leistungsmessungen in Baden-Württemberg aus unterschiedlichen Perspektiven beleuchtet werden soll. Dabei werden jährlich durchgeführte quantitative Befragungen mit qualitativen Studien gekoppelt. Die erste Erhebungsrunde 2006 bezog sich auf die erstmals verpflichtend durchgeführten Diagnose- und Vergleichsarbeiten am Ende des Schuljahres 2005/06. In diesem Artikel werden Ergebnisse der ersten Interviewstudie mit Lehrkräften berichtet. Parallel dazu wurden Lehrkräfte in Baden-Württemberg per Fragebogen postalisch zu Akzeptanz, Nutzung und schulinternem Umgang mit Vergleichsarbeitsrückmeldungen befragt (Maier, im Druck).

In beiden Teilstudien werden Selbstauskünfte von Lehrkräften zur Rezeption und Nutzung von Leistungsrückmeldungen erhoben und analysiert. Die methodischen Einschränkungen bisheriger Rezeptionsstudien müssen auch für das Vorgehen in dieser Studie in Kauf genommen werden: Verzerrungen durch sozial erwünschte Antworttendenzen oder unterschiedliche Referenzrahmen für die Beurteilung des eigenen Handelns. Der Vorteil einer qualitativen Interviewstudie ist allerdings, dass durch konkretes Nachfragen einerseits und durch die zum Teil recht ausführlichen Antworten der Lehrkräfte andererseits die Realität besser abgebildet werden

kann als in Fragebogenerhebungen. Durch die Anonymisierung der Befragung war es den Lehrkräften möglich, auch kritisch und weniger „sozial erwünscht“ über Vergleichsarbeiten und deren Nutzung zu reden.

Trotz methodischer Schwachstellen lässt sich die Fokussierung auf die Lehrerperspektive vor allem theoretisch gut begründen. Adressaten von Vergleichsarbeiten und Testrückmeldungen sind in erster Linie Lehrkräfte. Sie sollen zu einer datengeleiteten, professionellen Reflexion über und Planung von Unterricht angeregt werden (z. B. Modell von Helmke & Hosenfeld 2005). Es macht also Sinn, Effekte zentraler Tests zunächst einmal genau an dieser Stelle zu untersuchen. Die Wirkungen von Vergleichsarbeiten auf Unterricht und letztendlich auf den Wissenserwerb der Schüler sind dagegen wesentlich indirekter und können selbst mit großem forschungsmethodologischem Aufwand von weiteren Einflussgrößen kaum getrennt werden. Dies zeigte sich bei den bisherigen Untersuchungen von „washback“-Effekten auf Schülerleistungen und Unterricht. Die Studien können Effekte anderer Reformelemente nicht separieren (Firestone, Winter & Fitz 2000), beschränken sich auf die Untersuchung weniger Einzelfälle (Cheng 1999) oder sind nur sehr eingeschränkt ökologisch valide (Kellaghan, Madaus & Airasian 1982; Coe 1998).

3.1 Stichprobe

Die Zusammensetzung der Stichprobe geht aus Tabelle 1 hervor. An der Grundschulinterviewstudie nahmen Schulen aus dem Landkreis Göppingen teil. Aus einer offiziellen Schulliste wurden 14 Grundschulen zufällig gezogen. Eine Grundschule verweigerte aufgrund Überlastung die Teilnahme an der Befragung. Bei den Sekundarschulen war die Quote der Ablehnung einer Teilnahme an der Studie höher, sodass in einem größeren Gebiet (Region Stuttgart und Ostwürttemberg) weitere Haupt- und Realschulen zufällig ausgewählt und angeschrieben wurden. Aussagen zu den Diagnose- und Vergleichsarbeiten in Mathematik liegen insgesamt von 18 Grundschullehrkräften und 32 Sekundarschullehrkräften vor. Aufgrund der qualitativ-explorativen Vorgehensweise werden keine Aussagen über Schulformunterschiede in der Sekundarstufe angestrebt. Damit stellt auch die Unterrepräsentation der Gymnasiallehrkräfte keine Beeinträchtigung der Ergebnisse dar.

	Schulen	Lehrer (m/w)	Klassen mit Leistungsdaten	Interviews Mathematik
GS	13	20 (4/16)	19 Klassen ∅ 22,0 Schüler und 3,5 Migranten	18
HS	13	17	15 Klassen	15

		(7/10)	∅ 20,3 Schüler und. 8,0 Migranten	
RS	8	19 (5/14)	16 Klassen ∅ 26,8 Schüler und. 3,9 Migranten	13
GY	2	4 (3/1)	4 Klassen (keine weiteren Angaben vorhanden)	4

Tabelle 1: Stichprobe für die Interviewstudie 2006

Um die Diskrepanz zwischen Leistungsmessung bzw. Notengebung der Lehrkraft und dem Anspruchsniveau der Diagnose- und Vergleichsarbeit auch für die Klassen der an der Studie beteiligten Lehrkräfte beschreiben zu können, wurden im Interview auch individuelle Leistungsdaten erhoben. Für 18 Grundschulklassen (n = 392), 15 Hauptschulklassen (n = 300) und 16 Realschulklassen (n = 330) liegen individualisierte Mathematikleistungsdaten vor. Die Notenangaben der Gymnasiallehrkräfte werden aufgrund der kleinen Zahl nicht analysiert.

3.2 Interviews und Datenauswertung

Die Interviews wurden im Zeitraum Juli bis Oktober 2006 an den jeweiligen Schulen der befragten Lehrkräfte durchgeführt. Grundlage für das Interview war ein halbstrukturierter Leitfragebogen zu den Bereichen Rezeption, Reflexion, Akzeptanz und schulinterne Nutzung der Tests und Leistungsrückmeldungen. Zu jedem Bereich konnten vom Interviewer vertiefende Zusatzfragen gestellt werden. Ziel war es, mit dieser Form des halbstrukturierten, problemzentrierten Leitfadeninterviews eine aus Sicht des Befragten erschöpfende und angemessene Beschreibung des Untersuchungsgegenstandes zu ermöglichen (Witzel 1982). Um die Aussagen der Lehrkräfte zu den einzelnen Fragebereichen dennoch miteinander vergleichen zu können, musste deshalb auf die Auswertungstechnik der qualitativen Inhaltsanalyse zurückgegriffen werden (Mayring 2000). Dieses Verfahren eignet sich in besonderem Maße, um eine große Anzahl qualitativer Interviews auf bestimmte Fragestellungen hin zu analysieren und zu systematisieren (Maier 2005).

Hierzu wurde der Interviewtext in einzelne Aussagenelemente zerlegt. Die Aussagenelemente wurden paraphrasiert und den einzelnen Fragebereichen zugeordnet. In einer ersten konsensuellen Validierung kontrollierten die drei an der Auswertung beteiligten Personen die Richtigkeit der Zuordnung zum Fragenbereich. Anschließend wurden alle Aussagen innerhalb eines Fragenbereichs nach Ähnlichkeit, d. h. induktiv sortiert. Ähnliche Aussagen (Paraphrasen) wurden in einer Kategorie zusammengefasst und mit einem passenden Label bezeichnet. In einer zweiten konsensuellen Validierung wurden diese Kategorienzuordnungen noch einmal kontrolliert und bei Bedarf diskutiert und neu geordnet.

Nach Abschluss der Kategorienbildung konnten die Kategorienlabels innerhalb eines Fragenbereiches aufgelistet und quantifiziert werden. Dabei spielten zwei Informationen eine Rolle:

- Wie viele Aussagen wurden dieser Kategorie zugeordnet?
- Wie viele Lehrkräfte haben Aussagen zu dieser Kategorie beigesteuert?

Durch diese Form der Quantifizierung ist es möglich, die Relevanz bestimmter qualitativer Aussagen abzuschätzen. Aufgrund der Stichprobengröße und -auswahl dürfen diese Quantifizierungen allerdings nicht als zu generalisierende Werte missverstanden werden.

4 Ergebnisse

Zunächst einmal werden die Benotungsmaßstäbe der befragten Lehrkräfte mit den Testleistungen ihrer Klassen verglichen. Diese Analyse gibt erste Hinweise auf den von den Lehrkräften subjektiv wahrgenommenen Schwierigkeitsgrad der Vergleichsarbeiten. Anschließend werden die Ergebnisse der qualitativen Inhaltsanalyse getrennt nach Primar- und Sekundarstufe dargestellt.

4.1 Leistungsniveau der beteiligten Klassen

Die Mehrheit der an der Interviewstudie beteiligten Lehrkräfte hat von sehr einfachen Diagnose- und Vergleichsarbeiten im Vergleich zu den eigenen Leistungsmessungen berichtet. Diese Wahrnehmung deckt sich ebenfalls mit einer vom Landesinstitut für Schulentwicklung durchgeführten empirischen Analyse des Anspruchsniveaus der zentralen Tests. Hierzu wurden die landesweiten Punktedurchschnitte der Pilotierung im Jahr 2005 mit dem Pflichteinsatz im Jahr 2006 verglichen[3]. Dabei wurde festgestellt, dass erhebliche Differenzen im Leistungsniveau der getesteten Schüler vorliegen. Der Vergleich zeigte eine erfolgreichere Bearbeitung der Aufgaben im Pflichteinsatz im Vergleich zur Pilotierung ein Jahr zuvor. In der Grundschule betrifft dies die Diagnosearbeiten in den beiden Hauptfächern. In der Hauptschule und der Realschule sind vor allem die Vergleichsarbeiten in Mathematik wesentlich besser ausgefallen als bei der Pilotierung. Im Gymnasium waren es ebenfalls beide Hauptfächer Deutsch und Mathematik, in denen die Schüler im Pflichteinsatz bessere Ergebnisse erzielen konnten. Dies erklärt die von den Lehrkräften wahrgenommene und vielfach geäußerte „Einfachheit“ der Diagnose- und Vergleichsarbeiten.

---

[3] Landesinstitut für Schulentwicklung: Erfahrungen mit Pilotierung 2005 und Pflichteinsatz 2006. Online veröffentlicht unter http://lbsneu.schule-bw.de/entwicklung/dva/dva_docs/ Walldorf/Erfahrungen_Walldorf.pdf [08.06.2007]

Der auf Individualebene durchgeführte Vergleich zwischen Testnote und Jahresendnote belegt die tendenziell mildere Bewertung bzw. das geringere Anspruchsniveau der ersten verpflichtenden Diagnose- und Vergleichsarbeiten aus Sicht der an der Studie beteiligten Lehrer. Für die befragten Hauptschullehrkräfte ergibt sich allerdings eine Diskrepanz zwischen durchschnittlicher Zeugnisnote in Mathematik (3,2) und durchschnittlicher Vergleichsarbeitsnote (2,8). Lediglich in der Realschule stimmten Zeugnisnoten und Testnoten im Schnitt überein (jeweils 3,1). In der Grundschule ist kein direkter Notenvergleich möglich, da der Test nicht benotet werden darf und keine Umrechnungstabelle vorgegeben wurde.

Die vom Landesinstitut eingeräumte geringe Schwierigkeit der ersten verpflichtenden Diagnose- und Vergleichsarbeiten gibt zunächst einmal eine objektive Erklärung für die Notendiskrepanzen. Die mit Tests eng verknüpfte soziale Kontrollsituation in Lehrerkollegien muss allerdings auch in Rechnung gestellt werden. Lehrkräfte werden vermutlich in eigenen Klassenarbeiten höhere Ansprüche stellen und strenger beurteilen als in zentral gestellten Tests. Einerseits demonstrieren Lehrkräfte durch „harte Klassenarbeiten“ ihr subjektiv empfundenes, fachliches Anspruchsniveau. Andererseits wird jeder Lehrer an einem guten Abschneiden seiner eigenen Klasse bei einem Vergleichstest interessiert sein. Während sich bei Klassenarbeiten Schüler und Lehrer oft gegenüber stehen, bildet sich bei einem zentral gestellten Test viel eher eine gewisse „Schicksalsgemeinschaft“. In einigen internationalen Studien wurde dieser „Solidarisierungseffekt“ mit Schülern beschrieben. Lehrkräfte bereiten ihre Schüler auf die Tests vor, geben ihnen die ein oder andere Hilfestellung und bewerten bei vorhandenen Spielräumen eher zugunsten des Schülers (z. B.Stecher 2002).

4.3 Konsequenzen für den Mathematikunterricht

Die qualitative Inhaltsanalyse führte zu fünf induktiv gewonnenen Kategorien. Die Kategorien werden folgendermaßen bezeichnet und definiert:

- „Keine Veränderung/Bestätigung des Unterrichts“: Lehrkräfte ziehen keine nennenswerten Konsequenzen oder sehen die Testrückmeldung weitgehend als Bestätigung des eigenen Unterrichts an.
- „Bestimmte Aufgaben übernehmen/üben“: Es wird von einzelnen Testaufgaben berichtet, die zum Nachdenken angeregt haben und von der Lehrkraft eventuell in den zukünftigen Mathematikunterricht integriert oder verstärkt geübt werden.
- „Mehr Wiederholungen“: Die Diagnose- und Vergleichsarbeiten wurden durch Übungs- und Wiederholungseinheiten vorbereitet bzw. für den nächsten Durchgang werden entsprechende Wiederholungen fest eingeplant.

- „Lernmethodische Konsequenzen“: Die Tests gaben dem Lehrer Hinweise auf lernmethodische oder arbeitstechnische Schwächen der Schüler.
- „Allgemeine Konsequenzen“: In einer Art Restkategorie wurden abstrakte und eher wenig verbindlich erscheinende Aussagen, die sich auf den zukünftigen Mathematikunterricht beziehen, gesammelt.

Die Verteilung der extrahierten Nennungen auf die fünf Kategorien ist sehr ungleich und wird als Übersicht in Tabelle 2 dargestellt. Dabei werden jeweils zwei Werte angegeben. Die Anzahl der Lehrkräfte, die Aussagen zu der entsprechenden Kategorie beigesteuert haben, kann als grobes Maß für die Quantität interpretiert werden. Die Anzahl der kategorisierten Einzelnennungen dagegen weist auf einzelne Lehrkräfte hin, die in diesem Bereich sehr elaborierte Aussagen machen konnten und deshalb mehrfach vertreten sind.

Kategorie	Anzahl Nennungen (Lehrer/Aussagen)	
	Sekundarstufe I (Kl. 6) (n = 32)	Grundschule (Kl. 2) (n = 18)
Keine Veränderung/Bestätigung	15/19	9/16
Bestimmte Aufgaben übernehmen/ üben	8/9	7/11
Mehr Wiederholungseinheiten	6/8	–
Lernmethodische Konsequenzen	2/2	3/4
Allgemeine Konsequenzen	4/4	5/5

Tabelle 2: Kategorienübersicht und Häufigkeitsverteilung der Einzelnennungen

Ein Blick auf die Häufigkeiten in Tabelle 2 zeigt bereits das zentrale Ergebnis der qualitativen Auswertung. Die Hälfte der Grundschullehrkräfte als auch der Sekundarschullehrkräfte sieht aufgrund der Testergebnisse keinen nennenswerten Veränderungsbedarf und zieht damit auch keine Konsequenzen für den zukünftigen Mathematikunterricht. Wenn Konsequenzen gezogen werden, dann geht es um die verstärkte Beachtung bestimmter Aufgabenstellungen aus den Tests oder um die generelle Betonung von gezielten Wiederholungseinheiten zur Testvorbereitung. Der Kategorie „Mehr Wiederholungen“ wurde keine Nennung eines Grundschullehrers zugeordnet. Einige Grundschullehrkräfte sprachen zwar von Wiederholungen, die in Zukunft nötig sind, verbanden diese Aussage jedoch immer mit konkreten Aufgabenstellungen, die verstärkt zu üben sind.

In den folgenden Abschnitten werden die Kategorien genauer beschrieben. Zur besseren Veranschaulichung werden ebenfalls Lehrerzitate eingefügt. Die Auswahl der Lehrerzitate orientiert sich weitgehend am Kriterium der Verständlichkeit für diesen Ergebnisbericht. Einige Lehreraussagen sind zu lang oder nur im Zusammenhang mit anderen Interviewabschnitten verständlich. Ausgewählt wurden deshalb Zitate, mit denen sich der Inhalt der Kategorie möglichst klar erfassen lässt. Mehrere Lehreraussagen werden zitiert, wenn innerhalb einer Kategorie verschiedene Teilaspekte bedeutsam sind.

4.2.1 Sekundarstufe

Keine Veränderung bzw. Bestätigung des Unterrichts (15 Lehrkräfte; 19 Aussagen)

Rund die Hälfte der an Hauptschulen, Realschulen und Gymnasien befragten Lehrkräfte sah in den Ergebnissen der Mathematikvergleichsarbeit eine generelle Bestätigung des eigenen Unterrichts. Eine Ableitung von Schlussfolgerungen für den Unterricht wurde als nicht nötig erachtet. Zwar erwähnten einige dieser Lehrkräfte im weiteren Verlauf des Gesprächs doch noch die ein oder andere Konsequenz. Diese wurden allerdings immer als nebensächlich betrachtet bzw. als Ergebnis einer allgemeinen Reflexion über den eigenen Unterricht.

Es wurden keine Konsequenzen für den weiteren Unterricht gezogen, weil nach Meinung einiger Lehrkräfte die Vergleichsarbeitsrückmeldungen keine zusätzlichen Erkenntnisse gebracht haben: „Ich weiß als Klassenlehrer eigentlich ganz genau, wo es fehlt und wo es nicht fehlt. Da brauche ich jetzt keine Vergleichsarbeit. Ich weiß ganz genau, wo Lücken sind und wo nicht und wie ich fördern kann und wie nicht.“ (HS 02) Ebenfalls sehr häufig wurde das gute Gesamtergebnis der Klasse bei den ersten verpflichtenden Tests als Bestätigung des eigenen Unterrichts verbucht: „Und sonst brauche ich ja nicht irgendwelche Dinge machen. Wenn ich besser liege als der Schnitt, brauche ich ja nicht nachprüfen, wo kann ich noch besser.“ (HS 21)

Es gibt Lehrkräfte, die genauer begründen, warum sie die Vergleichsarbeit als Bestätigung des eigenen Vorgehens interpretieren können. Ein Hauptschullehrer sieht in den Rückmeldungen die Früchte seiner praktizierten „Denkerziehung“: „Ich mache schon immer solche Konzentrationsaufgaben oder Denkerziehung ist für mich etwas ganz Wichtiges. So wie ich in Deutsch eine rhetorische Erziehung mache, so mache ich in Mathematik eine Denkerziehung, gekoppelt mit einem Konzentrations- und Gedächtnistraining. Und insofern habe ich mich bestätigt gefühlt und habe auch mal eine Aufgabe beschmunzelt und habe auch gesagt, mhm, genau so ist es. Oder was Ähnliches hast du ja auch schon mal gemacht.“ (HS 08).

Ein Gymnasiallehrer betonte, dass nicht die zentrale Tests, sondern strukturelle Reformen wie die Verkürzung der Gymnasialzeit die eigentlichen Veränderungen

auslösen: „Ha, ich habe gesehen, dass mein Unterricht nicht so schlecht sein kann. Also es hat ja funktioniert. Nein, durch die Vergleichsarbeiten kamen keine Veränderungen. Die Veränderungen kamen vorher schon mit dem G8." (GY 02) Ein weiterer Gymnasiallehrer möchte strategisch vorgehen und erst die Vergleichsarbeiten 2007 abwarten, um zu sehen, wie er seine jetzigen Fünftklässler gezielt vorbereiten kann.

Mit der starken Abhängigkeit der Testergebnisse von der jeweiligen Klasse wird in einer Aussage begründet, dass generelle Veränderungen am Vorgehen im Mathematikunterricht nicht möglich sind: „Ich habe dieses Jahr wieder eine sechste Klasse und stelle fest, dass die Probleme an ganz anderen Stellen sitzen als letztes Jahr. Stellen, die letztes Jahr gut funktioniert haben und auch in der Vergleichsarbeit gut funktioniert haben, die machen meinen jetzigen Schülern Probleme. Und umgekehrt auch. Stellen, wo meine letzten Schüler Probleme hatten, klappen dieses Jahr besser. In der Planung habe ich keine Änderung gemacht. Ich habe das letztes Jahr auch schon so gemacht." (GY 04)

Bestimmte Aufgaben übernehmen (8 Lehrkräfte; 9 Aussagen)

Jede vierte Lehrkraft in der Sekundarstufe berichtet von bestimmten Testaufgaben, die für den eigenen Unterricht übernommen oder in Zukunft verstärkt geübt werden sollen. Ein Hauptschullehrer war von den Rechenaufgaben mit relativ einfachen, kleinen Zahlen beeindruckt: „Im Buch, vor allem in Mathe, gibt's halt Aufgaben, die haben relativ große Zahlen. Das ist natürlich schwierig für die Kinder. Die würde ich dann vielleicht eher zum Teil weglassen und mich wirklich auf einfache Grundaufgaben konzentrieren. Das, denke ich, hängt schon irgendwie ein bisschen mit der Vergleichsarbeit zusammen." (HS 02) Ein weiterer Hauptschullehrer zog Konsequenzen aus einer Aufgabe, in der die Schüler mithilfe eines Weg-Zeit-Diagramms den Verlauf einer Radtour erklären sollten: „Bei dieser einen Aufgabe Nummer 15. Das war diese Aufgabe. Das haben meine einfach nicht verstanden, dass da einfach eine Pause stattfindet. (…) Das würde ich dann noch öfters machen, solche Diagramme lesen." (HS 11)

Ein Realschullehrer möchte bei Textaufgaben in Zukunft die Fragen weglassen: „Was ich mir manchmal denke, (…) dass ich von Textaufgaben dann zum Beispiel die Fragen weglasse. Dass die Schüler selber schauen können." (RS 15) Und ein Gymnasiallehrer sieht sich durch die Vergleichsarbeit veranlasst, mehr Textaufgaben mit Problemlösecharakter zu stellen: „Vielleicht lege ich manchmal ein bisschen mehr Wert auf Textaufgaben oder solche Problemlöseaufgaben oder so was (…)." (GY 01)

Ein Realschullehrer betont, die Schüler mehr mit testähnlichen Aufgaben zu konfrontieren: „Dass die Schüler sich auch an diese Art und Weise gewöhnen, von

Blättern und Aufgaben. Denn die Aufgaben im Buch sind ja zum Teil doch anders gestellt.“ (RS 24) Ein Hauptschullehrer zieht ebenfalls die Konsequenz, mehr Multiple-Choice-Aufgaben zu nutzen: „Ich würde vielleicht die Aufgabentypen ein bisschen einbauen in Klassenarbeiten. Das wäre vielleicht noch eine Möglichkeit, dass man dieses Multiple-Choice-Verfahren, was ja eigentlich immer mehr im Kommen ist und war in der Vergangenheit, das würde ich eventuell mal einbauen.“ (HS 02)

Mehr Wiederholungseinheiten (6 Lehrkräfte; 8 Aussagen)

Eine weitere Kategorie ist die Nutzung der Vergleichsarbeiten zur Begründung umfangreicher Übungs- und Wiederholungseinheiten. Die nachfolgenden Aussagen beziehen sich zwar retrospektiv auf die durchgeführte Testvorbereitung, man kann allerdings mit großer Wahrscheinlichkeit davon ausgehen, dass die Lehrkräfte dies auch in Zukunft ähnlich handhaben werden. Beispielsweise ein Realschullehrer: „Das war mir ein willkommener Anlass. (…) Also ich habe kräftig mit denen geübt.“ (RS 21) Ebenso wird auch davon gesprochen, den selbst empfundenen „Druck“ an die Schüler weiterzugeben: „Also ich finde es immer ganz schön, man kann auf die Schüler ein bisschen größeren Druck ausüben. Also ich habe die explizit ja vorbereitet auf die Vergleichsarbeit. Ich hab noch mal drei, vier Wochen vorher alle wichtigen Dinge wiederholt.“ (RS 17)

Die zentralen Tests werden auch als willkommener Anlass für gezielte Übungen während der Schulferien angesehen: „Interessanterweise, ich habe an dem Elternabend (…) gesagt, ob sie was dagegen hätten, wenn ich den Kindern über die Ferien so Mathe-Arbeitsblätter gebe. Weil ich immer das Gefühl habe, die sitzen daheim, zumindest unsere Schüler, vor dem PC zwei Wochen lang in den Osterferien und dann kommen sie danach und sind total von der Rolle. Ja, und was sie davon halten. Und die Eltern, also alle die am Elternabend da waren, haben gesagt: Ja, machen Sie das bitte. Ich habe das jetzt noch mal nachgefragt. Ich werde es dieses Jahr auch so machen. Ich habe jetzt gerade zwei Arbeitsblätter ausgegeben. Die Schüler bruddeln natürlich. (…) Aber es ist trotzdem gut.“ (221)

Lernmethodische Konsequenzen (2 Lehrkräfte; 2 Aussagen)

Ein Hauptschullehrer hat sich vor allem über die vielen Leichtsinnsfehler seiner Schüler in der Vergleichsarbeit geärgert und möchte verstärkt auf genaues Rechnen und die richtige Verwendung von Maßeinheiten achten. Ein weiterer Kollege möchte das Lesen und Erstellen von Schaubildern und Diagrammen als grundlegende Kompetenz auch in anderen Fächern üben.

Allgemeine Konsequenzen (4 Lehrkräfte; 4 Aussagen)

In einer Art „Restkategorie“ wurden weitere Konsequenzen gesammelt. Beispielsweise die Schlussfolgerung, beim nächsten Durchgang „die Themen noch schneller durchzuziehen“, um gegen Schuljahresende alles abgedeckt zu haben. Ein Hauptschullehrer denkt an eine Verlagerung inhaltlicher Schwerpunkte: „Ich weiß, dass ich meine Schwerpunkte wahrscheinlich anders legen muss, zum Beispiel mehr auf Bruchrechnung und mehr auf Interpretation von Schaubildern.“ (HS 22)

Lediglich ein Hauptschullehrer möchte aufgrund der Vergleichsarbeiten in Zukunft mehr differenzieren und „Kärtchen oder Arbeitsblätter“ vorbereiten (HS 03). Zwei Hauptschullehrkräfte nehmen sich vor, mathematische Inhalte anschaulicher aufzubereiten, langsamer zu bearbeiten und „mehr mit Anschauungsmaterialien und fassbaren Beispielen“ zu arbeiten (HS 22). Ein Lehrer kommt zu dieser Schlussfolgerung, nachdem er eine Testaufgabe zum räumlichen Denken noch einmal im Unterricht aufgegriffen hatte: „Ja, viel anschaulicher, viel langsamer noch. Ich habe dann anschließend ja auch Würfel und Dinge basteln lassen. Es ist eine Katastrophe.“ (HS 13)

4.2.2 Grundschule

Keine Veränderungen bzw. Bestätigung des Unterrichts (9 Lehrkräfte; 16 Aussagen)

Analog zu den Ergebnissen in der Sekundarstufe sieht gut die Hälfte der befragten Grundschullehrkräfte keinerlei Anlass, den eigenen Mathematikunterricht aufgrund der Testrückmeldungen nennenswert zu verändern. Das relativ gute Abschneiden bestätigt die Grundschullehrer, einen guten, abwechslungsreichen und lehrplankonformen Unterricht zu machen. Es wird ebenfalls argumentiert, dass die Diagnosearbeiten durch die Schulbücher gut „abgedeckt“ sind und somit keine großartigen Umstellungen notwendig sind. Ein Zitat hierzu: „Das bestätigt eigentlich im Großen und Ganzen das, was man bisher so an Zielen verfolgt, also was Neues eigentlich weniger. Dass es jetzt neue Impulse setzt? Weniger.“ (GS 04)

Einige Grundschullehrkräfte werden konkreter und nennen Maßnahmen, auf die sie das gute Testergebnis zurückführen und die somit auch nicht weiter verändert oder verbessert werden müssen: „Die individuellen Fördermaßnahmen und (die) Differenzierung innerhalb des Unterrichts mache ich aufgrund jetzt der Diagnosearbeiten auch nicht anders wie vorher. Kinder, die Hilfe brauchen, fasse ich als Gruppe zusammen, während die anderen still weiterarbeiten.“ (GS 03) Oder ein bereits eingeschlagener Weg zur Veränderung des Unterrichts wird bestätigt: „Wir haben schon früh angefangen mit Wörterbucharbeit oder mit freierem Arbeiten im Mathematikunterricht oder weg von so automatisierten Aufgaben und mehr Denkaufgaben und das hat unseren Weg eigentlich bestätigt.“ (GS 16)

Immer wieder kann man zwischen den Zeilen der Interviewaussagen auch den Stolz der Lehrkräfte herauslesen, mit der eigenen Klasse relativ gut abgeschnitten zu haben: „Diese letzte Aufgabe hier. Das haben die wunderbar hingekriegt. Oder gerade die hier mit dieser Spiegelachse. Oder hier diese Uhrzeiten. Waren also fast keine Fehler. Also ein paar Kollegen haben gesagt: Au, das ist schwer. (…) Aber da haben wir halt kräftig geübt." (GS 18) An Veränderungen des Unterrichts zu denken, erübrigt sich nach dieser Interpretation der Testergebnisse.

Bestimmte Aufgaben übernehmen (7 Lehrkräfte; 11 Aussagen[4])

Jeder dritte befragte Grundschullehrer erwähnte bestimmte Aufgabentypen, die den Schülern Schwierigkeiten bereiteten und zukünftig auch im eigenen Mathematikunterricht verstärkt Beachtung finden sollen. Dabei handelt es sich vor allem um Aufgaben zum Zahlenstrahl, zu Uhrzeiten und Zeitspannen, zu Tabellen und zum Tauschen von Münzen. Auch die Probleme der Schüler im Umgang mit Sachsituationen und Textaufgaben werden reflektiert und als wichtige Schwerpunktsetzung für den eigenen Unterricht angesehen.

Lediglich ein Grundschullehrer spricht gezielt von „produktiven Übungsaufgaben und Knobelaufgaben", die in den zukünftigen Mathematikunterricht eingebaut werden sollen. Derselbe Lehrer entwickelte noch weitere Ideen: „Dann wirklich auch noch mehr so Rechengeschichten erzählen, Rechengeschichten erfinden zu Aufgaben. Also das habe ich aus dieser einen Rechenaufgabe abgeleitet. Also wirklich die Verbindung Rechengeschichte und die auch sprachlich formuliert zu einer Rechenaufgabe. Und umgekehrt. (…) Also das war auch so eine Erkenntnis aus der Diagnosearbeit für mich. (…) Eine Rechengeschichte haben und da im Prinzip eine Aufgabe dazu. Und das wirklich Schüler formulieren zu lassen." (GS 20)

Lernmethodische Konsequenzen (3 Lehrkräfte, 4 Aussagen)

Drei Grundschullehrkräfte ziehen lernmethodische Konsequenzen aus den Testergebnissen. Beispielsweise verstärkt darauf zu achten, dass Schüler schriftliche Arbeitsanweisungen selbständig lesen und umsetzen können. Bei einer weiteren Lehrkraft steht ebenfalls die Förderung des selbständigen Arbeitens im Mittelpunkt der Schlussfolgerungen: „Also vor allem bei Aufgabenstellungen, die die Selbständigkeit fördern. Das versuche ich jetzt viel stärker in meinem Unterricht oder stärker jetzt zu berücksichtigen, dass Kinder auch mal mit Aufgaben konfrontiert werden, wo selbständiges Lernen gefragt ist. Selber auf Lösungen mal kommen, also stärker ausprobieren, Lösungswege alleine versuchen zu finden oder auch in der

[4] Die Lehrkraft GS 20 hat allein mit fünf Aussagen zu dieser Kategorie beigetragen und muss insgesamt als „Ausnahme" betrachtet werden.

Gruppe. Also diese Selbständigkeit, das ist schon für mich ein Ziel, das ich stärker berücksichtigen will.“ (GS 17)

Ebenso wird daran gedacht, die gegliederte Darstellung von Aufgaben zu üben oder Schüler auf alternative Lösungsformen hinzuweisen: „Auch, ich sage mal, Schüler dazu anzuleiten - wobei das ich vorher, vor der Diagnosearbeit schon probiert habe - also so zeichnerische Lösungen. Ja, also wenn Kinder dann dasitzen und irgendwie jetzt den Text lesen, einfach zu sagen, versuche es mal zu zeichnen. Vielleicht kommst du dann drauf. Versuche mal, die Geschichte zu zeichnen.“ (GS 20)

Allgemeine Hinweise (5 Lehrkräfte, 5 Aussagen)

In einer Restkategorie wurden wiederum eher allgemeine Konsequenzen gesammelt. Beispielsweise wird an eine Verschiebung der geometrischen Themengebiete oder mehr Differenzierung gedacht: „Also im nächsten - es kommt dieses Schuljahr nicht zum Tragen, weil ich jetzt Einser habe - aber nächstes Schuljahr wird es schon zum Tragen kommen. Und ich werde wirklich dann schwerpunktmäßig differenzieren.“ (GS 19) Eine Lehrkraft möchte eigene Schwerpunkte über das Lehrbuch hinaus setzen, benennt diese aber nicht konkret. In einem Interview wird die Diagnosearbeit als allgemeiner Orientierungsrahmen gesehen: „Ich habe jetzt einen Eindruck, was hier gefordert wird und was wohl auch, ich gehe mal davon aus, am Ende der zweiten Klasse eben nach diesen Vergleichsarbeiten erwartet wird. Man kann gezielter da darauf eingehen.“ (GS 13)

5 Zusammenfassung und Diskussion

Zunächst einmal stellte sich die Frage, welche Konsequenzen aus den ersten baden-württembergischen Diagnose- und Vergleichsarbeiten für den zukünftigen Mathematikunterricht gezogen wurden. Es dominiert der Eindruck, dass die Mehrheit der befragten Lehrkräfte allein aufgrund der Testrückmeldung keine Schlussfolgerungen für den weiteren Mathematikunterricht zieht. Wenn von bestimmten Konsequenzen geredet wird, sind diese eher spärlich, randständig und könnten durchaus auch das Ergebnis einer nicht-testbasierten, kritischen Reflexion des eigenen Unterrichts sein.

Dennoch wurden in den Interviews einige Aufgaben erwähnt, die zu einer Reflexion über den eigenen Unterricht anregen konnten. Bei einer Aufgabe mussten auf der Grundlage eines Weg-Zeit-Diagramms Fragen zu einer Fahrradtour beantwortet werden. Diese Aufgabe ließ sich nur mit einem grundlegenden Verständnis des dargestellten funktionalen Zusammenhangs lösen. Die von einigen Lehrern beobachteten Verständnisdefizite der Schüler bei der Aufgabenbearbeitung konnten

zum Nachdenken über die verstärkte Einbindung von Diagrammen anregen. Genau in diesen Fällen findet eine Ableitung von Konsequenzen unter Rückbezug der Testergebnisse auf den vorausgegangenen Unterricht statt.

Diese Verknüpfung zwischen Rückmeldung und dem Aufbau mathematischer Konzepte ist allerdings die große Ausnahme in den analysierten Interviews. Wenn die Testergebnisse auf den vorausgegangenen Mathematikunterricht bezogen werden, dann erfolgt dies in der Regel in Form einer generellen Bestätigung der eigenen Arbeit. Gute Ergebnisse werden als Lob verbucht. Diese Form der Verstärkung des Lehrerhandelns durch Vergleichsarbeiten kann durchaus ambivalent betrachtet werden. Einerseits ist es gut, dass Lehrkräfte durch Tests eine Bestätigung erfahren, weil Lob im Berufsalltag viel zu selten vorkommt. Andererseits lenkt diese Form der pauschalen Verstärkung von den in den Tests enthaltenen Detailinformationen ab und führt somit nicht zu dem intendierten Lerneffekt durch Feedback (Hattie & Timperley 2007).

Gründe für diese mangelhafte Nutzung von Testrückmeldungen könnten nun einerseits bei den Vergleichsarbeiten und dem in Baden-Württemberg etablierten Rückmeldeformat gesucht werden. Andererseits könnte man nach Lehrermerkmalen fragen, die eine reichhaltigere Reflexion zentraler Leistungsrückmeldungen begünstigen. Beide Ursachenbereiche sollen vor dem Hintergrund der Befragungsergebnisse ausgelotet werden. Da es sich um eine explorative und auf ein Bundesland und einen bestimmten Testzeitpunkt beschränkte Datenerhebung handelt, sind die hier diskutierten Ergebnisse als weiterführende und noch zu belegende Hypothesen zu betrachten.

5.1 Merkmale der Tests und Rückmeldeformate

Bei einem Vergleich mit anderen Bundesländern fällt sehr schnell auf, dass in Baden-Württemberg Vergleichsarbeiten und Rückmeldeformate eingeführt wurden, die sich vor allem durch eine gewisse „Sparsamkeit“ auszeichnen. Die Pilotierungsstichprobe ist nicht repräsentativ und dennoch werden diese Ergebnisse als Referenzwerte zur Verfügung gestellt. Die Testergebnisse orientieren sich außerdem nicht an einem Kompetenzmodell. Andererseits wird von den Lehrern selbst verlangt, dass sie ihre eigenen Leistungsmessungen nicht an Sozialnormen, sondern an Lehrplanzielen orientieren.

Dagegen wird in Vergleichsarbeitsprojekten mit größerer und langjähriger Testerfahrung mittlerweile von gewissen Mindestanforderungen an Testaufgaben und einer Mindestmenge an statistischen Informationen ausgegangen. 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 (Büchter & Leuders 2005a). Im

Rahmen von VERA und den nordrhein-westfälischen Lernstandserhebungen wurden entsprechende Kriterien für Testaufgaben und Rückmeldungen entwickelt (Blum et al. 2005; Peek et al. 2006). Entlang dieser Anforderungen sollen nun mögliche Ursachen für die spärliche Nutzung der baden-württembergischen Testrückmeldungen lokalisiert werden.

Gute Testaufgaben sollten möglichst objektiv auswertbar sein und den Anforderungen einer eindimensionalen Rasch-Skalierung genügen. Dies schließt Aufgaben mit unterschiedlichen Lösungswegen oder Lösungen zunächst aus. Gerade aber komplexere Aufgaben regen zu einer mathematischen Erkundung an und repräsentieren anspruchsvolle mathematische Standards. Gute Testaufgaben sind somit eher weniger gute Lernaufgaben und können nur eingeschränkt zu einer Weiterentwicklung der Aufgabenkultur anregen. In den nordrhein-westfälischen Lernstandserhebungen geht man deshalb Kompromisse ein und verzichtet bei einem Teil der Aufgaben auf die psychometrische Qualität zugunsten des didaktischen Anregungspotenzials (Büchter & Leuders 2005b).

Wie wird dieser Konflikt nun bei den baden-württembergischen Vergleichsarbeiten gelöst? Für die Testitems werden einfache Aufgabenschwierigkeiten nach der klassischen Testtheorie berechnet. Auf eine Rasch-Skalierung und die Lokalisierung von Testitems innerhalb eines Kompetenzmodells wurde gänzlich verzichtet. Andererseits repräsentieren die Testaufgaben eine eher traditionelle Aufgabenkultur und nur an einigen Stellen lassen sich innovative Impulse erkennen. Die einzelnen Aufgaben werden in der Handreichung zwar den entsprechenden Standards zugeordnet, im Vordergrund steht allerdings die Überprüfung basaler Rechenalgorithmen und mathematischer Konzepte. Die Aufgaben zur Modellierung von Sachsituationen weichen kaum von traditionellen Sachaufgaben ab.

Die Lehrkräfte sollten ebenfalls bei einer gezielten Fehleranalyse unterstützt werden. Bei Multiple-Choice Items ist dies in der Regel nur schwer möglich, weil der Lösungsraum eingeschränkt wird und die Ratewahrscheinlichkeit auch bei vier Distraktoren keine zuverlässigen Rückschlüsse auf Denk- und Lösungswege der Schüler zulässt. Aber in den baden-württembergischen Vergleichsarbeiten findet man eine Reihe halboffener Aufgabenstellungen, die sich durchaus für eine individuelle Fehleranalyse eignen würden. In Nordrhein-Westfalen wurden hierzu qualitative Fehleranalysen in Pilotstudien durchgeführt und Fehlertypen identifiziert. Diese wurden den detaillierten Aufgabenbeschreibungen angehängt und können von Lehrkräften für eine vertiefte Analyse und eine gezielte Weiterarbeit genutzt werden (Büchter & Leuders 2005b). Auch diese Hinweise fehlen in den Handreichungen für die Diagnose- und Vergleichsarbeiten.

Weitere wichtige Bedingungen für die Innovationskraft zentraler Tests sind Praktikabilität, Relevanz und Verständlichkeit der Leistungsrückmeldungen (Peek et al.

2006). Als Standard gilt hier mittlerweile die Rückmeldung von Kompetenzprofilen und Aufgabenlösungshäufigkeiten auf Klassenebene. Die Klassenergebnisse können über „faire Vergleiche“ und unter Angabe von Konfidenzintervallen mit Parallelklassen oder anderen, vergleichbaren Schulen verglichen werden. Diese Daten sind aufgrund des Aggregationsniveaus hinreichend reliabel und ermöglichen der Lehrkraft eine Gesamtbeurteilung der durch den Unterricht aufgebauten mathematischen Kompetenzen vor dem Hintergrund der angebotenen sozialen und kriterialen Bezugsnormen (Büchter & Leuders 2005b; Lorenz 2005; Nachtigall & Kröhne 2006; Peek & Dobbelstein 2006).

Die baden-württembergischen Leistungsrückmeldungen verzichten dagegen sowohl auf den fairen Vergleich als auch auf die Rückmeldung von Kompetenzprofilen auf Klassenebene. In den Pilotstudien wurden Quartilsintervalle berechnet, die als Referenzrahmen für die klasseninterne Punkteverteilung dienen. Das heißt, dem Lehrer steht allenfalls eine sehr grobe soziale Bezugsnorm zur Verfügung, um die Leistungsfähigkeit seiner Schüler insgesamt und damit auch die Qualität seines Unterrichts einschätzen zu können. Für eine genauere Interpretation kann er auf die Aufgabenanalyse zurückgreifen. Aber auch hier werden keine Kompetenzbereiche ausgewiesen und der soziale Vergleich orientiert sich ebenfalls an der Pilotstichprobe. Der Informationsgehalt der Testrückmeldungen unterschreitet somit den üblichen Standard. Gerade die für Lehrkräfte interessanten und vielleicht auch „brisanten“ Vergleiche werden nicht geliefert.

Letztendlich sollte auch über die Form der geforderten Reflexion nachgedacht werden. Während in Baden-Württemberg und auch anderen Bundesländern recht vage die Idee einer kollegialen Diskussion über Testergebnisse angemahnt wird, wurde im Projekt Check 5 des Schweizer Kantons Aargau ein ganz anderer Weg gewählt. Die Lehrkräfte sind zunächst einmal verpflichtet, die aus den Testrückmeldungen abgeleiteten Konsequenzen schriftlich zu fixieren. Ebenso müssen die Testergebnisse und die Schlussfolgerungen in einer frei wählbaren Form mit den Schülern und Eltern besprochen werden (Tresch 2007). Beides sind Handlungen, die sich in einem größeren Maße überprüfen lassen und somit zu einer höheren Verbindlichkeit beitragen.

Aufgrund der Lehreräußerungen ist auch erkennbar, dass Vergleichsarbeiten im Unterricht vorbereitet werden. Diese Reaktion ist verständlich und aufgrund bisheriger Befunde auch erwartbar. Allerdings zeigt sich genau an dieser Stelle auch der Funktionskonflikt, in dem zentrale Tests insgesamt stehen. Wenn Sie den Unterricht verändern sollen, ist eine Vorbereitung erwünscht. Allerdings kann man dann nicht mehr von einer objektiven Prüfung von Leistungen im Sinne der empirischen Bildungsforschung reden. Diese Funktionskonflikte oder auch die Funktionsfülle

wird in ähnlicher Weise auch in anderen Bundesländern deutlich kritisiert (Peek et al. 2006; Sill & Sikora 2007).

Ein weiterer Beleg für die eher geringe Objektivität der Vergleichsarbeiten ist die von einigen Lehrkräften beschriebene Abhängigkeit der Testergebnisse von Schülerjahrgängen. Somit müsste auch in BW noch viel deutlicher herausgestellt werden, dass Vergleichsarbeiten vor allem eine individuelle Rückmeldung für die einzelne Lehrkraft sind. Sobald aber die Ergebnisse nach außen gehen, und sei es auch nur an die Schulleitung, muss eine Leistungsmessung möglichst objektiv und fair sein, um als solche von Lehrkräften akzeptiert zu werden.

5.2 Merkmale der Lehrkräfte

In dieser Studie wurden die Merkmale der befragten Lehrkräfte nicht systematisch erfasst. Dennoch soll zu heuristischen Zwecken auf eine interessante, singuläre Beobachtung bei der Auswertung der Interviews eingegangen werden. Ein Grundschullehrer mit ca. fünf Dienstjahren (GS20) fiel durch eine sehr detaillierte Reflexion der Testergebnisse sowie eine im Vergleich zu Kollegen umfangreiche Nutzung der Rückmeldungen besonders auf. Im gesamten Interviewtext wurden deshalb weitere Hinweise gesucht, mit denen sich ein Bild dieser Lehrkraft zeichnen lässt.

Auffallend war zunächst seine sehr kritische und offen geäußerte Einstellung zu den Diagnosearbeiten. Der Lehrer bezweifelte das Aufwand-Nutzen-Verhältnis und dies obwohl er bei weitem die meisten Konsequenzen für den eigenen Unterricht ziehen konnte. Er kritisierte vor allem den Zeitpunkt der Tests, weil damit keine gezielte Individualförderung durch dieselbe Lehrkraft mehr möglich ist. Während des Interviews war immer wieder diese starke Schülerfokussierung in der Argumentation erkennbar.

Den individualdiagnostischen Gehalt der Rückmeldungen beurteilte der Lehrer ebenfalls als sehr gering. Die Daten wurden allenfalls als Bestätigung der eigenen Beobachtungen verbucht. Dies zeigte sich auch daran, dass für lernschwache Schüler der Klasse bereits Fördermaßnahmen bestanden und im Interview detailliert über Stärken und Schwächen einzelner Schüler berichtet werden konnte. Die vorliegenden Leistungsergebnisse der Diagnosearbeit waren hierzu lediglich ein Gesprächsimpuls und wurden vor dem Hintergrund der individualdiagnostischen Expertise erklärt.

Von besonderer Relevanz war die Tatsache, dass der Grundschullehrer die Testaufgaben immer wieder mit den im Bildungsplan vorgegebenen Kompetenzen sowie mit Lehrbuchaufgaben, die „nach dem neuen Bildungsplan vorgehen“ in Verbindung bringen konnte. Er äußerte sogar den Eindruck, dass die Diagnosearbeit bewusst Inhalte und Kompetenzen des neuen Bildungsplans abprüfen möchte. Auf

die Frage nach der Lehrplanvalidität einzelner Testaufgaben argumentierte er mit der neuen, kompetenzorientierten Begrifflichkeit, die charakteristisch für die mathematikdidaktischen Standards im Bildungsplan der Hauptschule ist. Einige Aufgaben prüften aus seiner Sicht ganz klar „mathematisches Verständnis" ab.

Wiederum kritisch sah er die Abhängigkeit des Testergebnisses vom eingesetzten Schulbuch, was auch andere befragte Grundschullehrkräfte bemängelten. Ebenfalls geht aus einigen Aussagen eine kritische Beschäftigung mit der Gültigkeit und Reichweite zentraler Testergebnisse hervor. Er betonte, dass man mit landesweiten Diagnosearbeiten nur einen kleinen Teil der im Mathematikunterricht zu erwerbenden Kompetenzen abprüfen kann. Dennoch nahm er die ganze Sache sehr ernst und die Diagnosearbeit in Mathematik war trotz kritischer Grundeinstellung für ihn Anlass genug, noch einmal genauer in den Bildungsplan zu schauen: „Also ich habe mir dann nochmal – muss ich ganz ehrlich sagen – den Bildungsplan erst mal so richtig genau angekuckt nach der Diagnosearbeit (lacht). Was jetzt Mathe betrifft, so ganz genau durchgelesen." (GS20)

Diese Skizze einer einzelnen Lehrkraft lässt sich natürlich nicht verallgemeinern, gibt aber deutliche Hinweise auf professionelle Kompetenzen, die als Bedingungen für eine sinnvolle Nutzung zentraler Leistungsrückmeldungen in Frage kommen könnten: klare Schülerorientierung, diagnostisches und fachdidaktisches Wissen, Bereitschaft zur Weiterbildung und die Fähigkeit zur Selbstkritik. Dies deutet auf eine noch zu lösende Grundproblematik bei der Entwicklung und Einführung zentraler Tests hin. Es gibt bisher noch kein schlüssiges theoretisches Konzept, wie eine Veränderung der Unterrichtspraxis bzw. des Lehrerlernens im Anschluss an Vergleichsarbeiten stattfinden soll. Die reine Durchsicht und Interpretation von Ergebniswerten impliziert sehr viele mögliche Reaktionen, und eine gezielte Verbesserung des Unterrichts ist ein mögliches Zufallsprodukt, kann jedoch auch komplett ausbleiben. Eine sinnvolle Lösung scheint vor allem die systematische Verbindung mit Lehrerfortbildungen, speziell der Fachkollegien in den Einzelschulen zu sein (Sill & Sikora 2007).

Auch für Blum et al. (2005) macht die Einführung zentraler Tests zur Qualitätssicherung nur dann Sinn, wenn Lehrerfortbildungen oder Unterrichtsentwicklungskonzepte als flankierende Maßnahmen ein kohärentes Gesamtkonzept ergeben. In Thüringen beispielsweise werden die landesweiten Kompetenztests durch ein ständiges Fortbildungsangebot ergänzt[5]. Dabei steht nicht nur das Lesen und Interpretieren statistischer Daten im Mittelpunkt. Vielmehr geht es um die Einbindung der Testrückmeldungen in eine datengestützte, schulinterne Evaluation. Dieses Modell könnte auch in Baden-Württemberg die Weiterentwicklung zentraler Tests anre-

[5] www.kompetenztest.de [20.01.2008]

gen, wenn jährliche Vergleichsarbeiten mehr als ein weitgehend von der eigenen Unterrichtspraxis isoliertes Ritual sein sollen.

Literatur

Amrein, A. L. & Berliner, D. C. (2002): High-stakes testing, uncertainty and student learning. Education Policy Analysis Archives, 10(18).

Amrein, A. L. & Berliner, D. C. (2003): The effects of highstakes testing on student motivation and learning. Educational Leadership, 60(5), 32–38.

Baeriswyl, F.; Wandeler, C.; Trautwein, U. & Oswald, K. (2006): Leistungstest, Offenheit von Bildungsgängen und obligatorische Beratung der Eltern. Zeitschrift für Erziehungswissenschaft, 9(3), 371–392.

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(4), 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, Weinheim: Beltz, 125–148.

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(4), 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(4), 324–334.

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

Coe, R. (1998): Feedback, Value Added and Teachers' Attitudes: Models, Theories and Experiments. Unpublished PhD thesis, University of Durham.

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(1), 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. New York, München, Berlin: Waxmann, 19–40.

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

Herman, J. L. (2004): The Effects of Testing on Instruction. In: S. H. Fuhrman & R. F. Elmore (Hrsg.): Redesigning Accountability Systems for Education. New York, London: Teachers College Press, 141–166.

Kellaghan, T.; Madaus, G.F. & Airasian, P.W. (1982): The Effects of Standardized Testing. London: Kluwen, Nijholf Publishing.

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

Klieme, E.; Avenarius, H.; Blum, W.; Döbrich, P.; Gruber, H.; Prenzel, M.; Reiss, K.; Riquarts, K.; Rost, J.; Tenorth, H.-E. & Vollmer, H. J. (2003): Zur Entwicklung nationaler Bildungsstandards – Eine Expertise. Berlin: BMBF.

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

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

Koretz, D.; Stecher, B.; Klein, S. & McCaffrey, D. (1994): The Vermont portfolio assessment program: Findings and implications. Educational Measurement: Issues and Practice, 13(3), 5–16.

Kultusministerkonferenz (2003): Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss. Beschluss der Kultusministerkonferenz vom 4.12.2003.

Lind, G. (2004): Jenseits von PISA – Für eine neue Evaluationskultur. In: Pädagogische Hochschule Schwäbisch Gmünd, Institut für Schulentwicklung (Hrsg): Standards, Evaluation und neue Methoden. Reaktionen auf die PISA-Studie. Baltmannsweiler: Schneider Verlag Hohengehren, 1–7

Lorenz, J. H. (2005): Zentrale Lernstandsmessung in der Primarstufe: Vergleichsarbeiten Klasse 4 (VERA) in sieben Bundesländern (Central student assessment in primary schools: comparative tests for grade 4 in seven federal states of Germany). Zentralblatt für Didaktik der Mathematik, 37(4), 317–324.

Maier, U. (2005): Lehrerverhalten und Emotionen aus Sicht von Schülern. In: P. Mayring & M. Gläser-Zikuda (Hrsg.): Die Praxis der Qualitativen Inhaltsanalyse. Weinheim: Beltz-UTB, 190–206.

Maier, U. (im Druck): Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. Akzeptiertes Manuskript. Zeitschrift für Pädagogik.

Mayring, P. (2000): Qualitative Inhaltsanalyse. In: U. Flick, E. v. Kardorff & I. Steinke (Hrsg.): Qualitative Forschung. Ein Handbuch, Reinbek: Rowohlt, 468–475.

McDonnell, L. M. & Choisser, C. (1997): Testing and Teaching: Local Implementation of New State Assessments. CSE Technical Report No. 442. Los Angeles: University of California, Center for the Study of Evaluation (CRESST).

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. [abgerufen am 9.1.2007 unter www.lehrmittelverlag.ch/downloads/dateien/Evaluation%20Klassencockpit.pdf].

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

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. New York, München, Berlin: Waxmann, 59–74.

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. New York, München, Berlin: Waxmann, 41–58.

Peek, R. & Nilshorn, I. (2004): Schulrückmeldungen von Schulleistungsstudien am Beispiel des QuaSUM-Projektes: Zwei Untersuchungen zur Wirksamkeit der Schulforschung in Brandenburg, Heft 3, Ministerium für Bildung, Jugend und Sport des Landes Brandenburg.

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

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: F. Eder, A. Gastager & F. Hofmann (Hrsg.). Qualität durch Standards, Münster: Waxmann, 219–233.

Popham, W. J. (1987): The merits of measurement-driven instruction. In: Phi Delta Kappa, 68, S. 679–682.

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

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

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

Stecher, B. M. & Mitchell, K.J. (1995): Portfolio Driven Reform; Vermont Teachers' Understanding of Mathematical Problem Solving. CSE Technical Report 400. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

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

Witzel, A. (1982): Verfahren der qualitativen Sozialforschung. Überblick und Alternativen. Frankfurt a. M.: Campus.

Wottawa, H. & Thierau, H. (1998): Lehrbuch Evaluation. 2. vollst. überarb. Auflage. Bern: Huber.

Anschrift des Autors

Dr. Uwe Maier
Pädagogische Hochschule Schwäbisch Gmünd
Oberbettringerstraße 200, 73525 Schwäbisch Gmünd
uwe.maier@ph-gmuend.de

Eingang Manuskript: 13.09.2007 (überarbeitetes Manuskript: 22.01.2008)