Maier, Uwe; Rauin, Udo

Vergleichsarbeiten - Hilfe zur Unterrichtsentwicklung? Zentrale Lernstandserhebungen aus Sicht baden-württembergischer Lehrkräfte

Die Deutsche Schule 98 (2006) 4, S. 403-421

Uwe Maier und Udo Rauin

Vergleichsarbeiten – Hilfe zur Unterrichtsentwicklung?

Zentrale Lernstandserhebungen aus Sicht baden-württembergischer Lehrkräfte

Vergleichsarbeiten[1] werden häufig als das notwendige Pendant zu Bildungsstandards betrachtet, denn sie ermöglichen die *Überprüfung der Standards auf den einzelnen Ebenen des Schulsystems* und sollen Informationen für die ergebnisorientierte Schul- und Unterrichtsentwicklung generieren (Ditton 2002; Helmke 2003, Klieme et al. 2003). Mit der verbindlichen Einführung nationaler Bildungsstandards in Deutschland wurden gleichzeitig von den Bundesländern erste Pilotprojekte zur Einführung zentraler und flächendeckender Lernstandserhebungen gestartet. In Baden-Württemberg wurden erstmals gegen Ende des Schuljahres 2005/06 verpflichtende Vergleichsarbeiten in den Klassen 6 und 8 der weiterführenden Schulen und Diagnosearbeiten in der Grundschulklasse 2 durchgeführt. Aber werden Vergleichsarbeiten von den Lehrkräften akzeptiert und in der erwarteten Weise genutzt?

Bereits in den Jahren 2003 und 2004 hatten Schulen bzw. einzelne Lehrkräfte die Möglichkeit, freiwillig an den Pilotstudien zur Erprobung der zentralen Lernstandsmessungen teilzunehmen. Durch die Möglichkeit der freiwilligen Teilnahme entstand eine quasi-experimentelle Situation. Eine Gruppe von Lehrkräften konnte bereits im Vorfeld der verpflichtenden Einführung praktische Erfahrungen mit dem neuen Evaluationsinstrument machen. Hierzu wurde in zwei repräsentativen Lehrerbefragungen die Akzeptanz und der pädagogische Nutzen zentraler Lernstandserhebungen aus Lehrerperspektive in Abhängigkeit der bisher gemachten Testerfahrungen erhoben. Die Befragung der Hauptschul-, Realschul- und Gymnasiallehrer war Teil einer längsschnittlich angelegten Studie zur Einführung von Bildungsstandards in Baden-Württemberg (2003 – 2006). Die Grundschullehrkräfte wurden separat im Herbst 2005 befragt. Im Zentrum beider Teilstudien stand jeweils die Frage, ob Diagnose- und Vergleichsarbeiten aus Sicht der Lehrkräfte pädagogisch relevante Daten zur Verfügung stellen können oder die Kritik an einem neuen „Kontrollinstrument" überwiegt. Dies erlaubt Rückschlüsse auf die derzeitige Situation, in der Vergleichsarbeiten für alle Lehrkräfte verbindlich eingeführt sind.

1 Für zentrale Lernstandserhebungen im Rahmen der Überprüfung von Bildungsstandards werden je nach Bundesland unterschiedliche Begriffe verwendet (v. Ackeren & Bellenberg 2004). In diesem Text soll der in der Literatur am häufigsten verwendete Begriff „Vergleichsarbeiten" für flächendeckende, zentral gestellte Leistungserhebungen stehen.

1. Problemstellung

Die Entwicklung nationaler Bildungsstandards und die Erprobung bzw. verpflichtende Einführung zentraler, flächendeckender Lernstandsmessungen in sämtlichen Bundesländern sind nicht nur die bildungspolitische Konsequenz aus enttäuschenden Leistungsergebnissen im internationalen Vergleich, sondern stehen unter dem generellen Vorzeichen eines *Paradigmenwechsels in der Schulsystemsteuerung*, der bereits vor Jahrzehnten eingeleitet wurde (vgl. Klieme 2004; Böttcher 2004; v. Ackeren 2004). Böttcher (2002) fasst diese Entwicklungen zusammen und spricht von zwei schultheoretischen Problembereichen, die den Rahmen einer Diskussion über Standards und Vergleichsarbeiten bilden: Es geht einerseits um eine Qualitätsverbesserung durch die Stärkung der Einzelschule (Schulautonomie) und andererseits um die Kontrolle der Resultate durch externe Instanzen. Bildungsstandards und zentrale Lernstandsmessungen sind das notwendige Bindeglied zwischen Dezentralisierung und Rezentralisierung.

Nachdem die Diskussion über Bildungsstandards ihren Höhepunkt überschritten hat und Modelle der Outputsteuerung nach und nach in das deutsche Schulsystem Einzug halten, stellt sich für die Bildungsforschung die Frage nach der Wirksamkeit bzw. nach Wirkmechanismen von Standards und standardbezogenen Evaluationen. Die gesamte Diskussion zu diesem Thema wird jedoch im wesentlichen von ungeprüften Annahmen und bildungspolitischen Wunschvorstellungen dominiert. Klieme (2004, 633) fasst die Situation folgendermaßen zusammen: „Die zentralen Probleme dürften in der für Deutschland noch weitgehend ungeklärten Verknüpfung zwischen Standards, Schulentwicklung, Schulevaluation und Rechenschaftslegung liegen." Um die sehr weitreichende Thematik eingrenzen zu können, werden zunächst allgemeine und international diskutierte Wirkungsmodelle für Standards und evaluative Rückmeldesysteme vorgestellt. Danach wird speziell auf Modelle und Rezeptionsstudien zu den hier interessierenden Vergleichsarbeiten in Deutschland eingegangen.

1.1 Wirkungsmodelle für Standards und Leistungsrückmeldungen

Die Einbettung von Bildungsstandards und Vergleichsarbeiten in die sehr breit angelegten *schultheoretischen Diskussionslinien* über Qualitätssicherung, Schulentwicklung, Schulautonomie und Schulsystemsteuerung erschwert eine gezielte Auswahl von Wirkungsmodellen. Je nach theoretischem Kontext variieren diese deutlich, operieren mit einer unterschiedlichen Semantik und sind mehr oder weniger empirisch fundiert. Böttcher (2004) beispielsweise kritisiert die bildungspolitische und administrative Argumentationslogik, die mit einem empirisch nicht haltbaren kybernetischen Regelkreismodell operiert. Bildungsstandards im Sinne von Leistungserwartungen stellen einen Sollwert (Input) dar, der nach erfolgtem Unterricht (Prozess) mittels zentraler Lernstandsmessungen mit dem Istwert (Outcome) verglichen wird. Diese Information wird an die Einzelschule bzw. die Lehrkraft zurückgemeldet und soll dort zusammen mit den Standards als „erweiterte Inputinformation" das Unterrichten als zentralen Prozess neu justieren und somit zu einer ergebnisorientierten Unterrichtsentwicklung führen. Die Feedbackschleife des Regelkreislaufs zweifelt Böttcher allerdings vehement an. Für ihn ist das System Schule in hohem Maße „untersteuert", wie zum Beispiel auch die Lehrplan-

wirksamkeitsforschung deutlich zeigen konnte (Rauin, Tillmann & Vollstädt 1996; Vollstädt et al. 1999; Künzli 1999).

Komplexer und dem Gegenstand dadurch angemessener ist das Modell für Evaluation und Qualitätssicherung bei Ditton (2002). Es integriert die Ebenenstruktur des Bildungssystems in den bereits beschriebenen Regelkreislauf. Das Schulsystem schafft bestimmte Voraussetzungen bzw. Standards für schulische Bildungsprozesse. Hierzu gehören strukturelle, finanzielle und personelle Rahmenbedingungen aber auch vorgegebene Bildungsziele (intendiertes Curriculum). Diese Standards beeinflussen das implementierte Curriculum an Schulen (institutionelle Ebene) und im Unterricht (Interaktionsebene). Eine Überprüfung (assessment) bezieht sich dann auf die kurz- oder langfristigen Bildungswirkungen (erreichtes Curriculum). Von accountability bzw. evaluationszentrierter Kontrolle spricht Ditton, wenn sich die Überprüfung von Bildungswirkungen auf Standards (intendiertes Curriculum) bzw. die schulinternen Prozesse (implementiertes Curriculum) regulativ auswirken kann.

O'Day (2002; 2004) entwickelte auf der Grundlage umfangreicher Forschungsbefunde über die Wirkung von *school accountability systems*[2] in den USA ein theoretisches Rahmenmodell, das auf systemtheoretische Prämissen zurückgreift. Auch wenn die US-amerikanischen *accountability systems* größtenteils andere Zielsetzungen haben als die hierzulande diskutierten und teilweise implementierten Rückmeldesysteme, ist die theoretische Modellierung von O'Day interessant, weil dort speziell nach den Auswirkungen von Leistungsrückmeldungen für die Einzelschulentwicklung gefragt wird. Schulen bzw. Lehrer lernen, indem sie nur die Informationen aufnehmen, die für zukünftiges Handeln effektiv erscheinen. Außerdem müssen Informationen richtig interpretiert werden, um Entwicklungen in Gang setzen zu können. Dies ist in komplexen Systemen nicht immer der Fall. Einschränkungen (*constraints*) ergeben sich durch die individuelle Wissensbasis, soziale konstruierte Überzeugungen (*belief systems*) und die Komplexität der Interaktionen, die das System definieren. Aus diesen Überlegungen folgert O'Day ein Rahmenmodell für die Wirkung von Bildungsstandards. Standards und Tests sind dann für die Verbesserung von Lernen und Unterricht von Nutzen, wenn sie Informationen bereitstellen, die auf Unterricht und Lernen aufmerksam machen. Als weitere Bedingung muss auf Individual- und Systemebene Wissen entwickelt werden können, das eine valide Interpretation der Informationen unterstützt. Und nicht zuletzt müssen Rückmeldesysteme auch Informationen für eine bessere Verteilung der zur Verfügung stehenden Ressourcen generieren können.

Die empirischen Befunde zur Prüfung der Wirkung von *accountability systems* auf Schülerleistung werden allerdings auch von O'Day (2004) als widersprüchlich bezeichnet. Es scheint sich jedoch immer wieder zu bestätigen, dass Schulen höchst unterschiedlich auf die neuen Systeme der Rechenschaftslegung reagieren und die zurückgemeldeten Informationen auf ihre je eigene Art und Weise nutzen (DeBray/Parson/Woodworth 2001). Gut situierte Schulen mit einer bevorzugten Schülerschaft reagieren bereitwilliger und stimmiger auf die Anforderungen, die sich aus einem *test-based accountability system* ergeben.

---

2 O'Day unterscheidet davon so genannte „marktorientierte" accountability systems, die lediglich die Schulwahlfreiheit der Eltern stützen sollen.

Jedoch selbst bei Schulen mit anfänglich schlechten Schülerleistungen gab es ganz unterschiedliche Entwicklungen. Diese konnten vor allem auf schulinterne Bedingungen wie z.B. einer gemeinsamen Verantwortlichkeit (*internal accountability*) zurückgeführt werden (Abelmann/Elmore 1999).

Die europäische Diskussion über Standards, Evaluationssysteme und Schulentwicklung wird vor allem von England und den Niederlanden aus geführt. Untersucht wurden Rückmeldesysteme, die zumindest potenziell Unterricht und Schule verändern können und damit den in Deutschland diskutierten Vergleichsarbeiten recht nahe kommen. Scheerens, Glas und Thomas (2003) unterscheiden hierzu verschiedene Typen empirischer Untersuchungen zur Messung und Rückmeldung von Schülerleistungen. Internationale und nationale Leistungsstudien mit stichprobenbasiertem Vorgehen und Multi-Matrix-Desing dienen zur Überprüfung und Sicherung von Mindeststandards in den Kernfächern *(system monitoring)*. Rückschlüsse auf einzelschulische Entwicklungen sind mit diesen Studien nicht möglich. Dagegen intendieren regelmäßig durchgeführte und nicht veröffentlichte Erhebungen schülerbezogener Leistungsdaten ganz gezielt eine ergebnisorientierte Unterrichtsentwicklung und interne Evaluation an Schulen. In der internationalen Diskussion wird dieser Typ von Leistungsrückmeldungen an Einzelschulen als *school performance feedback systems* (SPFS) bezeichnet und folgendermaßen definiert: „information systems external to schools that provide them with confidential information on their performance and functioning as a basis for school self-evaluation" (Visscher/Coe 2003, 322).

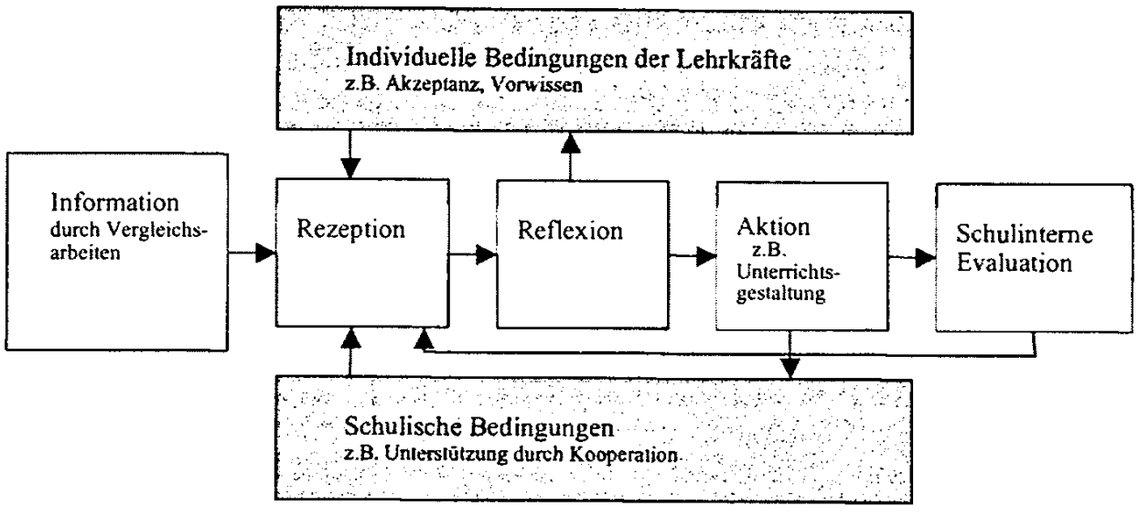
Die bisherige Forschungslage zu SPFS ist ebenfalls komplex und nicht eindeutig (Coe 2002). Die Unterschiedlichkeit der Feedbacksysteme und die Komplexität des Untersuchungsgegenstandes sowie die unterschiedlichen Anlagen der Studien erschweren allerdings eine Generalisierung bisheriger Ergebnisse. Dennoch schlagen Visscher und Coe (2003) nach einer Durchsicht einschlägiger internationaler Befunde ein Rahmenmodell vor, das dem von O'Day (2004) sehr nahe kommt. In den ersten Bereichen dieses Modells werden administrative Vorgaben (Entwicklung und Zielsetzung des Systems, Merkmale der Daten und der Implementation) sowie die schulischen Kontextfaktoren als wichtige Bedingungen für die Nutzbarmachung von Leistungsdaten beschrieben. Der Dreh- und Angelpunkt des Modells sind allerdings die unterschiedlichen Nutzungsmöglichkeiten und -strategien von Leistungsdaten auf der Schul- und Lehrerebene. Hierbei beziehen sie sich vor allem auf die Unterscheidung zwischen instrumenteller, konzeptueller und symbolischer Nutzung nach Rossi und Freeman (1993). Bei der instrumentellen Nutzung werden aufgrund der zur Verfügung stehenden Leistungsinformationen Entscheidungen getroffen. Beeinflusst die Rückmeldung lediglich das Denken der schulischen Entscheidungsträger, liegt eine konzeptuelle Nutzung vor. Werden dagegen die Feedback-Informationen ausschließlich selektiv genutzt, um den eigenen, bereits feststehenden Standpunkt argumentativ zu stützen, sprechen Rossi und Freeman von symbolischer Nutzung.

1.2 Vergleichsarbeiten und Rückmeldestudien in Deutschland

Im deutschsprachigen Raum scheint sich ein Typ standardbezogener Leistungsevaluation herauszukristallisieren, der sich deutlich von den *accountability systems* und *high-stakes tests* im anglo-amerikanischen Raum unterschei-

det und eher den SPFS zugeordnet werden kann. Primäres Ziel bisheriger Vergleichsarbeiten ist die Unterstützung der internen Schulevaluation. Die postulierten Wirkungen von Vergleichsarbeiten und Ergebnisrückmeldungen sind dennoch vielfältig und hängen von der Art des Tests, der Datenaufbereitung und der schulischen Implementation ab (z.B. Schrader/Helmke 2004). Gleichzeitig gibt es eine Reihe von mahnenden Stimmen, die vor übereilten Hoffnungen und überzogenen Forderungen warnen (z.B. Terhart 2002; Oelkers 2003). Rolff (2001) bezweifelt grundsätzlich die Verwendbarkeit rein deskriptiver Daten, die keine Erklärungskraft für Schulpraktiker besitzen und somit auch keine Veränderungswirkung entfalten können. Helmke (2003; 2004) beantwortet dagegen die Frage nach der pädagogischen Nutzbarmachung von Vergleichsarbeiten auf positive Weise. Er sieht einen möglichen Nutzen für die Elternberatung, die Standard- und Qualitätssicherung an Schulen, pädagogische Interventionen und die zielgerichtete Allokation von Fördermaßnahmen. Vergleichsarbeiten wird ein großes Potenzial für Schul- und Unterrichtsentwicklung zugeschrieben (Arnold 2002; Bonsen/von der Gathen 2004). Dabei werden faire Vergleiche zwischen Klassen und Schulen mithilfe adjustierter Daten als unverzichtbar angesehen (Nachtigall/Kröhne 2006).

Abbildung 1: Vereinfachtes Zyklenmodell nach Helmke & Hosenfeld (2005)



Im deutschsprachigen Raum wurde eine theoretische Modellierung postulierter Wirkungen von Standards und Vergleichsarbeiten auf Unterricht- und Schulentwicklungsprozesse von Helmke und Hosenfeld (2005) vorgelegt. Das Modell basiert ebenfalls auf dem bereits skizzierten kybernetischen Steuerungskreislauf und hat gewisse Parallelen zum systemtheoretisch orientierten Rahmenmodell von O'Day (2004). Die unterschiedlichen Informationen, die Vergleichsarbeiten liefern (linke Seite des Modells), müssen von den Lehrkräften rezipiert und reflektiert werden, bevor es zu konkreten Handlungen kommt. Der gesamte Prozess wird durch individuelle Voraussetzungen (Akzeptanz, Motivation, Vorwissen, professionelles Selbstverständnis) und schulische Bedingungen moderiert. Der Kreis schließt sich jedoch erst dann, wenn die getroffenen Maßnahmen innerhalb der Schule wieder überprüft werden (schulinterne Evaluation, rechte Seite des Modells) und sich in veränderten Aktionen (z.B. in der Unterrichtsqualität) oder in veränderten individuellen oder schulischen Bedingungen niederschlagen. Natürlich könnte man einwenden, dass eine Selbstreflexion der individuellen

Lehrkräfte oder der Konferenzen auch ohne Informationen durch Vergleichsarbeiten möglich wäre. Die Befürworter würden jedoch argumentieren, dass ein gut ausgebautes System von Vergleichsarbeiten (vgl. z.B. das mittlerweile breit erprobte Klassencockpit in der Schweiz; http://www.klassencockpit.ch) erst den Hintergrund liefert, um selbstreferentielle, nur auf die Norm der eigenen Klasse bezogene Interpretationen von Schülerleistungen zu überwinden.

Wenn man Rückmeldestudien im Rahmen von *large-scale assessments* ausklammert und die auf Länderebene verpflichtenden Vergleichsarbeiten betrachtet, findet man kaum empirische Studien, die den schulinternen Umgang mit zentralen Tests zum Gegenstand haben. In der Regel handelt es sich um Begleit- oder Evaluationsstudien mit explorativem Charakter, um die Handhabung und Interpretation von Vergleichsarbeiten zu optimieren. Bei VERA beispielsweise wurden 1510 Grundschullehrkräfte zum Umgang mit Vergleichsarbeiten und den Auswirkungen auf Unterrichtsentwicklungsmaßnahmen online befragt (Groß Opfhoff/Koch/Hosenfeld/Helmke 2006). Die Bedeutung der Fähigkeitsniveaus und die Unterschiede zwischen den Inhaltsgebieten eines Faches wurden in den Lehrerkollegien intensiv diskutiert. Dagegen wurde weniger über die sozialen Vergleichsmöglichkeiten geredet. Auch bezüglich der durch Vergleichsarbeiten angeregten Kooperation ziehen die Autoren ein positives Fazit. Eine überwiegende Mehrheit der Lehrkräfte gibt an, dass Diskussionen mit KollegInnen über Ursachen und Konsequenzen stattgefunden haben. Die Befunde sind allerdings nur schwer zu interpretieren, weil unklar bleibt in welchem Maße die zentrale Lernstandserhebung als Ursache identifiziert werden kann. Unklar bleibt ebenfalls, ob tatsächlich Änderungen stattgefunden haben und welches Ausmaß diese Änderungen hatten.

Für die Lernstandserhebungen in NRW in der Sekundarstufe I (Peek/Dobbelstein 2006) wurden den Schulen Anregungen zum Umgang mit den Ergebnissen auf unterschiedlichen Ebenen gegeben. Außerdem sind die Schulen dazu verpflichtet worden, in der Schulkonferenz über die Ergebnisse, deren Interpretation und daraus gezogene Konsequenzen zu berichten. Die Rückmeldung der Lehrkräfte an das Landesinstitut wurden bisher noch nicht systematisch evaluiert. Peek und Dobbelstein (2006) leiten aus dem bisherigen Lehrerfeedback weiterhin zu beachtende Dimensionen ab: Akzeptanz; Kompatibilität mit eigenen fachdidaktischen Vorstellungen; faire Vergleiche; innerschulische Kooperation. Auch die Thüringer Kompetenztests werden einer prozessbegleitenden Evaluation unterzogen. Dabei spielt die intensive Kooperation mit Lehrkräften eine wichtige Rolle. Es geht vor allem um die ständige Verbesserung und Anpassung der Rückmeldeformate mit dem Ziel der besseren Nutzbarkeit für Diagnose und Unterrichtsentwicklung (Nachtigall/Jantowski 2004).

Es fällt auf, dass es bisher keine unabhängigen Studien zur Abschätzung von Wirkungen auf der Schul- und Unterrichtsebene gibt. Die hier zitierten Projekte evaluieren ihre eigene Vorgehensweise, um die Tests selbst bzw. den Einsatz und die Interpretation der Tests zu optimieren.

2. Fragestellung

Die Einführung verpflichtender Lernstandsmessungen ist Teil der baden-württembergischen Bildungsplanreform 2004. Die turnusmäßig anstehende Lehrplanrevision wurde genutzt, um neben einem Bündel an schul- und unter-

richtsreformerischen Maßnahmen[3] allgemein verbindliche Bildungsstandards einzuführen, die mit den nationalen KMK-Standards allerdings wenig gemeinsam haben. Die flächendeckende Einführung von Diagnose- und Vergleichsarbeiten erfolgte dann zeitverzögert für das Schuljahr 2005/06 und wurde vom Landesinstitut für Schulentwicklung vorbereitet. Bereits in den Jahren 2003 und 2004 wurden erste Pilotstudien zur Erprobung der Tests durchgeführt und gleichzeitig interessierten Lehrkräften und Schulen via Internet zur freien Verfügung gestellt (Tabelle 1). Die Durchführung und Auswertung dieser freiwilligen Diagnose- und Vergleichsarbeiten erfolgte in Eigenregie mithilfe der vom Landesinstitut vorbereiteten Auswertungstabellen und Handreichungen. Die Schulen bzw. Lehrkräfte konnten ebenfalls selbst entscheiden, ob sie den Test als Klassenarbeit werten und wie sie die Reflexion der Ergebnisse gestalten. Die Einbindung der Diagnose- und Vergleichsarbeiten in eine schulinterne Evaluationskultur gilt als erklärtes Ziel, wird jedoch in den bisher vorliegenden Papieren nicht weitergehend erläutert.

Tabelle 1: Bisher durchgeführte Diagnose- und Vergleichsarbeiten

		2003	2004
Grundschule	Klasse 2: Klasse 3:	D, M D, M	D, M
Hauptschule	Klasse 5 Klasse 6	D, M	 D, E
Realschule	Klasse 6: Klasse 8	D, M, E	 D, M
Gymnasium (G9)	Klasse 6 Klasse 8:	 D, M, E, F	E, M

Die bisher durchgeführten Diagnose- und Vergleichsarbeiten wurden von erfahrenen Lehrkräften zusammengestellt und entsprechen „zentralen Klassenarbeiten". Eine Orientierung der Tests in der Pilotierungsphase an Kompetenzmodellen ist nicht erkennbar, bzw. wurde nicht berichtet. Den Lehrkräften wurden in einem zusätzlichen Dokument die nicht adjustierten Landesmittelwerte sowie die Interquartilbereiche mitgeteilt. Um eine klassenbezogene Aufgabenanalyse durchführen zu können, wurden zusätzlich noch die in der Pilotierungsstudie ermittelten Aufgabenschwierigkeiten berichtet. Weitere Vergleichswerte standen den Lehrkräften nicht zur Verfügung.

Daten dieser freiwilligen Erprobung von Diagnose- und Vergleichsarbeiten in den Jahren 2003 und 2004 werden genutzt, um die Sichtweisen von Lehrkräften mit und ohne Testerfahrung kontrastierend gegenüberzustellen. Dabei wird von folgender Überlegung ausgegangen: Lehrerinnen und Lehrer ohne Testerfahrungen haben gewisse Vorstellungen über zentrale Lernstandsmessungen und deren pädagogischen Nutzen. Diese Vorstellungen entsprechen der kollektiven Ausgangslage einer Lehrerschaft, die ohne verpflichtende, zentrale

3 Fächerverbünde, Schulcurriculum, Kontingentstundentafeln (vgl. www.bildung-staerkt-den-menschen.de)

Tests beruflich sozialisiert wurde. Ein gewisser Teil der Lehrkräfte macht nun Erfahrungen mit Diagnose- und Vergleichsarbeiten und kann die bisherigen Vorstellungen und Meinungen korrigieren. Der Vergleich zwischen beiden Gruppen ermöglicht somit eine erste Abschätzung der Dynamik, die durch die Einführung von Standards und zentralen Tests in Gang kommen könnte.

Die vorgestellten Wirkungsmodelle von Standards und zentralen Leistungstests (Visscher/Coe 2003; O'Day 2004; Böttcher 2004; Helmke/Hosenfeld 2005) betonen allesamt die zentrale Funktion der durch Lehrkräfte interpretierten Informationen von Leistungsrückmeldungen. Weder die Standards, noch die Outcome-Messung allein können im System Schule etwas bewirken. Auch die zurückgemeldeten Leistungsdaten und -vergleiche bleiben folgenlos, wenn sie von den Akteuren vor Ort nicht mit pädagogischer Bedeutung aufgeladen werden können. An dieser Stelle entscheidet sich, ob Bildungsstandards und zentrale Tests die gewünschten Folgen nach sich ziehen oder ob eher mit unerwünschten Nebenwirkungen gerechnet werden muss. Wünschenswert wäre eine Überprüfung des gesamten Modells, vor allem die Beschäftigung mit der Verarbeitung von Informationen und den daraus abgeleiteten Aktionen auf Schul- und Unterrichtsebene. Dazu liegen uns aber noch keine Daten vor. Die hier vorliegende Studie beschäftigt sich deshalb nur mit einem Teilbereich des Modells von Helmke und Hosenfeld (vgl. Abb. 1), nämlich mit der Rezeption. Sie soll die Sichtweise von Lehrkräften auf Vergleichsarbeiten während der Erprobungs- und Pilotierungsphase in Baden-Württemberg in Abhängigkeit von Vorerfahrungen und schulischen Bedingungen prüfen.
(1.) Wie werden Chancen und Risiken von Diagnose- und Vergleichsarbeiten in Abhängigkeit der bisherigen Testerfahrung eingeschätzt?
(2.) Lassen sich schulformspezifische Unterschiede in der Akzeptanz zentraler Leistungsmessungen feststellen?
(3.) Gibt es schulische Kontextfaktoren, die die Akzeptanz von zentralen Leistungsmessungen beeinflussen (Schulgröße, Einzugsgebiet, Migrantenanteil, Ganztagesschule)?

3. Stichproben und Instrumente

Grundlage für die Ergebnisdarstellung sind zwei Lehrerbefragungen, die im Rahmen einer längsschnittlich angelegten Studie zur Rezeption der Bildungsplanreform 2004 in Baden-Württemberg durchgeführt wurden. Die Befragung in der Sekundarstufe I wurde im Frühjahr 2005 durchgeführt und richtete sich an Hauptschul-, Realschul- und Gymnasiallehrkräfte der Klassen 5 und 6. Der Fragebogen umfasste die gesamte Palette der mit der Bildungsplanreform 2004 neu eingeführten Maßnahmen. An dieser Stelle werden lediglich die Daten zu den Vergleichsarbeiten berichtet. Die Grundschulbefragung erhebt Erfahrungen und Einstellungen von Grundschullehrkräften der Klassen 3 und 4 zu den Diagnosearbeiten und wurde im Herbst 2005 durchgeführt.

3.1.1 Stichprobenbeschreibung

Für die Lehrerbefragung in der Sekundarstufe I wurde aus einer Liste aller weiterführenden Schulen in Baden-Württemberg per Zufall jede zweite Hauptschule, jede zweite Realschule und jedes zweite Gymnasium gezogen. Die Schulleitungen wurden angeschrieben und gebeten, die Fragebögen an Lehrkräfte weiter-

zureichen, die in den Jahrgangsstufen 5 und 6 die Fächer Deutsch und Mathematik unterrichten. Für die Grundschulbefragung wurden 11 Land- bzw. Stadtkreise mit unterschiedlichen gymnasialen Übergangsquoten ausgewählt, um eine regional ausgewogene Stichprobe zu erhalten[4]. In diesen Kreisen wurden alle 495 Grundschulen angeschrieben und die Schulleitungen gebeten, die beiliegenden Fragebögen an Lehrkräfte der Klassenstufen 3 und 4 zu verteilen.

Tabelle 2: Stichprobe und Rücklauf

	Anzahl angeschriebener Schulen	Rücklauf Fragebögen	Prozentsatz Lehrerinnen
Grundschule	495	275	79,9 %
Hauptschule	590	605	67,1 %
Realschule	254	405	64,9 %
Gymnasium	247	284	54,6 %
gesamt	1091	1294	66,5 %

Der Rücklauf ist insgesamt zufriedenstellend, wenn man das zweistufige Verteilungsverfahren in Rechnung stellt (vgl. Tabelle 2). Eine exakte Rücklaufquote lässt sich nicht angeben, weil die Grundgesamtheit, d.h. die Anzahl der in Frage kommenden Lehrkräfte in den Jahrgangsstufen 5 und 6 nur grob geschätzt werden könnte. Vor allem an Hauptschulen ist unklar, ob eine oder zwei Lehrkräfte die Kernfächer unterrichten. In den Grund-, Haupt- und Realschulstichproben überwiegen die Lehrerinnen, in der Gymnasialstichprobe ist das Geschlechterverhältnis ausgewogen.

Die Altersverteilung spiegelt die langjährigen Lehrereinstellungszyklen in den einzelnen Schularten wider. Die Gruppe der 51- bis 60-Jährigen dominiert in allen Schularten die Kollegien (Hauptschulen: 34,0%; Realschulen: 39,8%; Gymnasien: 40,6%). Studienzeitbedingt ist die Gruppe der unter 30-Jährigen an Gymnasien (9,3%) wesentlich kleiner als an Real- und Hauptschulen (17,9% bzw. 21,8%).

Der Anteil der Lehrkräfte in der Sekundarstufe I mit Testerfahrung ist wesentlich geringer als in der Grundschule. Je nach Schulart differiert die freiwillige Teilnahme an den Vergleichsarbeiten in der Sekundarstufe I erheblich (vgl. Tabelle 3). An den Realschulen ist der Anteil am höchsten. In der Gymnasialstichprobe gibt nur jede fünfte Lehrkraft eine freiwillige Erprobung der zentralen Vergleichsarbeiten an. An den Hauptschulen ist die Testerfahrung ähnlich hoch wie an den Realschulen.

Tabelle 3: Testerfahrung der Lehrkräfte in der Sekundarstufe I

	Anteil Lehrkräfte mit Testerfahrung	davon ...	
		... freiwillig	... verpflichtend
Grundschule	61,6 %	59 %	2,1 %
Hauptschule	41,8 %	37,5 %	4,3 %
Realschule	48,3 %	45,3 %	3,0 %
Gymnasium	23,5 %	19,9 %	3,6 %

4 Datengrundlage: Statistisches Landesamt Baden-Württemberg

Da selbst in der Grundschulstichprobe über ein Drittel aller Lehrkräfte zum Zeitpunkt der Befragung noch keine Testerfahrung besitzen, lässt sich eine stabile Gegenüberstellung der Gruppen mit und ohne Testerfahrung in allen Schulformen durchführen

3.1.2 Instrumente

Die allgemeine Akzeptanz zentraler Leistungsmessungen wurde mit einer Skala von Ditton und Merz (2000) gemessen. Diese besteht aus sieben positiv und negativ gepolten Einzelitems, die den allgemeinen Nutzen zentraler Tests auf Schulebene beschreiben. Als Bezugspunkt dient jeweils der Begriff „zentrale, landesweite Testuntersuchungen", d.h. die Skala bezieht sich nicht direkt auf die neu eingeführten Diagnose- und Vergleichsarbeiten. (Z.B.: „Zentrale, landesweite Testuntersuchungen sind für die Arbeit der Schulen sehr wichtig." / „Zentrale, landesweite Testuntersuchungen nützen für meine eigentliche Arbeit als Lehrer wenig."). Die interne Konsistenz der Skala ist hoch (alpha = .88).

Der pädagogische Nutzen von Diagnose- und Vergleichsarbeiten wurde mit 6 positiv und negativ formulierten Items abgefragt. Eine explorative Hauptkomponentenanalyse mit den Daten der Sekundarlehrerbefragung führte zu den zwei Faktoren „förderdiagnostischer Nutzen von Vergleichsarbeiten" (F1) und „Risiken von Vergleichsarbeiten" (F2) (Tabelle 4). Mit den Daten der Grundschulbefragung konnte diese Faktorenstruktur nicht reproduziert werden, sodass für diesen Teil der Studie nur Ergebnisse auf Itemebene berichtet werden können.

Die internen Konsistenzen der neu gewonnenen Skalen „förderdiagnostischer Nutzen von Vergleichsarbeiten" und „Risiken von Vergleichsarbeiten" sind mit alpha-Werten von .75 und .70 zufriedenstellend. Die Interkorrelationen zwischen allen drei verwendeten Skalen sind signifikant ($p<0.01$) und liegen zwischen .52 und .59.

Tabelle 4: Faktorenstruktur der Items zum pädagogischen Nutzen von Vergleichsarbeiten (Befragung Sekundarstufe I)

Item	F1	F2
VA gut für Beratungsgespräche mit Eltern	,82	
Zur Planung von Fördermaßnahmen gut	,81	
VA guter Anhaltspunkt für Leistung einzelner Schüler	,70	-,42
Üben nur noch für den Test		,80
VA unnütz, weil Leistung vom Umfeld und Fähigkeiten abhängig		,76
Klassenarbeiten besser als VA		,71

Anmerkung: Hauptkomponentenanalyse mit Varimax und Vorgabe von zwei Faktoren; 65,5% erklärte Gesamtvarianz; Faktorladungen unter .30 nicht angezeigt

In der Grundschulstichprobe wurden die Lehrkräfte nach Schulgröße, Ganztagsbetreuung und Migrantenanteil (Prozentkategorien) gefragt. Als Maß der Schulgröße liegt die Anzahl der Parallelklassen vor. Da nur 10 der 274 befragten Grundschullehrkräfte angeben, dass sie an einer Ganztagsgrundschule unterrichten, wurden hierzu keine Untergruppenvergleiche gerechnet. Für die Be-

fragung der Lehrkräfte in der Sekundarstufe I liegen Daten zur Schulgröße, dem Einzugsgebiet der Schule (ländlich vs. städtisch) und dem Ganztagsangebot vor. Der Migrantenanteil wurde hier nicht erfragt.

4. Ergebnisse

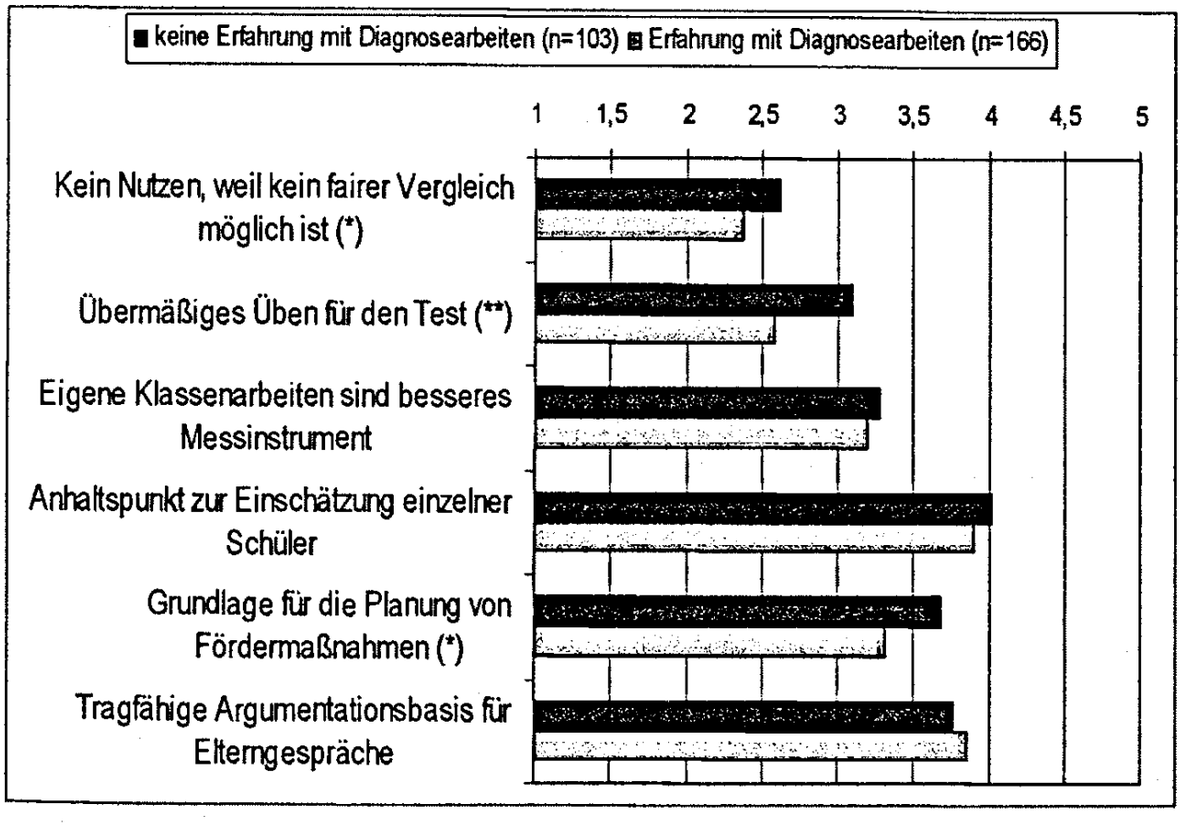
Zunächst werden Einstellungsunterschiede in Abhängigkeit der Testerfahrung geprüft und für beide Befragungen nach Schularten getrennt beschrieben. Im zweiten Teil der Ergebnisdarstellung wird der Einfluss externer Kontextfaktoren auf die Akzeptanz zentraler Tests geklärt. Aufgrund der oben beschriebenen Faktorenanalysen und Konsistenzprüfungen werden die Daten der Sekundarlehrerbefragung auf Skalenebenen analysiert. Bei den Daten der Grundschullehrerbefragung wird lediglich die „allgemeine Akzeptanz zentraler Tests" als Skala in die Analyse mit einbezogen. Die Bereiche „förderdiagnostischer Nutzen" und „Risiken von Diagnosearbeiten" werden auf Itemebene analysiert.

4.1 Akzeptanz zentraler Tests in Abhängigkeit von Testerfahrung und Schulart

4.1.1 Diagnosearbeiten in der Grundschule

Zentrale Lernstandserhebungen werden an Grundschulen im Durchschnitt eher befürwortet als abgelehnt. Die Skala „allgemeine Akzeptanz zentraler Tests" hat einen Mittelwert von 3,27. Grundschullehrkräfte, die bereits eine Diagnosearbeit durchgeführt haben, erreichen einen leicht höheren durchschnittlichen

Abbildung 2: Nutzen und Risiken von Diagnosearbeiten mit und ohne Testerfahrung (Einzelitems)



Anmerkung: 1 bedeutet Ablehnung des Items, 5 bedeutet Zustimmung zum Item.

Akzeptanzwert von 3,35, Lehrkräften ohne sind etwas neutraler (3,15). Diese Differenz war jedoch nicht signifikant (t-Test; p=0.062).

Die durchschnittlichen Einschätzungen der Items zu Nutzen und Risiken von Diagnosearbeiten in Abhängigkeit von der bisherigen Testerfahrung werden in Abbildung 2 dargestellt. Bei den Items zur kritischen Einschätzung der Diagnosearbeiten zeigt sich ein differenziertes Bild. Die Grundschullehrkräfte teilen die Kritik nicht, dass zentrale Tests überflüssig seien, weil sie keine fairen Vergleiche ermöglichen würden. Wenn Testerfahrungen vorhanden sind, wird dieses kritische Item noch stärker abgelehnt.

Die größte Differenz zwischen testerfahrenen und testunerfahrenen Grundschullehrkräften ergibt sich bei der Frage, ob durch Diagnosearbeiten ein übermäßiges Üben für den Test stattfindet (*teaching to the test*). Während testunerfahrene Lehrkräfte im Durchschnitt hier noch unentschieden bzw. in ihrer Meinung gespalten sind (höchste Standardabweichung von 1,29), wird dieses Item von Lehrkräften mit Testerfahrung mehrheitlich abgelehnt.

Etwas kritischer sind die Grundschullehrkräfte, wenn es um den Vergleich von Diagnosearbeiten mit den eigenen Klassenarbeiten geht. Von einer Mehrheit der befragten Lehrerinnen und Lehrer werden die eigenen Leistungsmessungen als zuverlässiger eingeschätzt. Wenn man allerdings bedenkt, dass die klasseninterne Leistungsmessung wesentlich häufiger stattfindet, sich direkt an den unterrichteten Inhalten orientiert und von den Lehrkräften selbst gestaltet wird, hätte man hier einen noch höheren Zustimmungswert erwarten können.

Der förderdiagnostische Nutzen der Diagnosearbeiten wird überwiegend positiv beurteilt. Unabhängig von der bisherigen Testerfahrung bewerten Grundschullehrkräfte die Diagnosearbeiten als guten Anhaltspunkt zur Einschätzung einzelner Schüler und als tragfähige Argumentationsbasis für Elterngespräche. Diese beiden Items erhalten insgesamt die höchsten durchschnittlichen Zustimmungswerte. Auch die Ableitung von geeigneten Fördermaßnahmen scheint aus Sicht der Grundschullehrerinnen und -lehrer möglich zu sein. Die Einschätzung der Lehrpersonen mit Testerfahrung ist hier jedoch verhaltener. Möglicherweise wurden ihre Erwartungen durch die Praxis der Vergleichsarbeiten enttäuscht.

Vergleichsarbeiten in der Sekundarstufe I

Der durchschnittliche Wert der Sekundarstufenlehrer auf der Skala „allgemeine Akzeptanz zentraler Tests" ist ebenfalls leicht positiv und damit vergleichbar mit den Akzeptanzwerten der Grundschullehrkräfte. Auch Lehrkräfte in der Sekundarstufe I äußern mit Testerfahrung eine signifikant höhere allgemeine Akzeptanz als Lehrkräfte ohne Erfahrung mit Vergleichsarbeiten (Tabelle 5).

Der förderdiagnostische Nutzen wird ebenfalls relativ hoch eingeschätzt. Auch hier haben Lehrkräfte mit Testerfahrung eine durchschnittlich günstigere Meinung. Die Risiken von Vergleichsarbeiten werden von Lehrkräften ohne Testerfahrung noch recht hoch eingeschätzt, während die Gruppe mit Testerfahrung hier etwas günstiger urteilt. Die signifikant höheren Standardabweichung bei beiden Skalen weisen allerdings darauf hin, dass die Meinungen bei Lehrkräften mit Testerfahrung weiter auseinander gehen (Levene-Test). Wir konnten diese Beobachtung auch auf Schulebene verfolgen. Sobald an den Schu-

Tabelle 5: Einschätzung der Vergleichsarbeiten in Abhängigkeit der Testerfahrung (Mittelwert und Standardabweichung)

	alle Lehrer (n=1274)	Mit Erfahrung (n=508)	Ohne Erfahrung (n=758)	t-Test	Levene-Test
Allgemeine Akzeptanz zentraler Tests	3,19 (0,87)	3,31 (0,88)	3,10 (0,85)	p = 0.000	n.s.
Förderdiagnostischer Nutzen von VA	3,35 (0,87)	3,43 (0,92)	3,30 (0,82)	p = 0.006	p = 0,006
Risiken von Vergleichsarbeiten	3,14 (0,91)	2,99 (0,98)	3,24 (0.85)	p = 0.000	p = 0,002

Anmerkung zur Skalierung: 1 ... Ablehnung / 5 ... Zustimmung.

len mit Vergleichsarbeiten gearbeitet wird, scheint es zu einer Polarisierung zwischen Befürwortern und Gegnern zu kommen.

Um mögliche Effekte der Schulart in Abhängigkeit der Testerfahrung zu prüfen, wurde eine zweifaktorielle Varianzanalyse mit den festen Faktoren Schulart und Testerfahrung gerechnet (Tabelle 6). Bei sämtlichen Skalen sind die Erfahrungseffekte signifikant und weisen in die bereits oben diskutierte Richtung. Ein statistisch bedeutsamer Unterschied zwischen den Schularten findet sich jedoch lediglich bei der Einschätzung des förderdiagnostischen Nutzens von Vergleichsarbeiten. Dieser wird an Hauptschulen insgesamt höher eingeschätzt als an Realschulen oder Gymnasien. Ob sich diese Differenz auf mögliche Unterschiede zwischen den eingesetzten Vergleichsarbeiten oder auf eine größeres Interesse an förderdiagnostischen Informationen bei Hauptschullehrkräften zurückführen lässt, bleibt hinter den quantitativen Daten verborgen.

Tabelle 6: Einschätzung der Vergleichsarbeiten nach Schulart und Testerfahrung (Mittelwert und Standardabweichung)

	VA-Erfahrung	HS (n=588)	RS (n=395)	GY (n=278)	Schulformeffekt	Erfahrungseffekt
Allgemeine Akzeptanz zentraler Tests	ohne	3,14 (0,83)	3,14 (0,84)	3,02 (0,88)	n.s.	p = 0.000
	mit	3,28 (0,89)	3,38 (0,87)	3,23 (0,85)		
Förderdiagn. Nutzen von Vergleichsarbeiten	ohne	3,45 (0,77)	3.18 (0,84)	3.16 (0,84)	p = 0.000	p = 0.045
	mit	3,55 (0,91)	3,35 (0,89)	3,21 (0,98)		
Risiken von Vergleichsarbeiten	ohne	3,24 (0,89)	3,21 (0,83)	3,26 (0,81)	n.s.	p = 0.000
	mit	3,03 (0,99)	2,90 (0,99)	3,10 (0,86)		

Anmerkung: Zweifaktorielle Varianzanalyse mit den festen Faktoren „Schulart" und „Erfahrung mit VA"; Keine signifikanten Interaktionseffekte; Skalierung: 1 ... Ablehnung / 5 ... Zustimmung.

Auch die insgesamt sehr positiven Einschätzung des förderdiagnostischen Nutzens von Vergleichsarbeiten durch Hauptschullehrer mit Testerfahrung zeigt gleichzeitig den bereits angesprochenen Effekt der Polarisierung. Dies gilt in ähnlicher Weise für Realschullehrkräfte. Die Befürchtung von Risiken nimmt auch hier mit Testerfahrung signifikant ab, die Varianz erhöht sich jedoch ebenfalls signifikant (Levene-Test auf Homogenität der Varianzen: p=0,017).

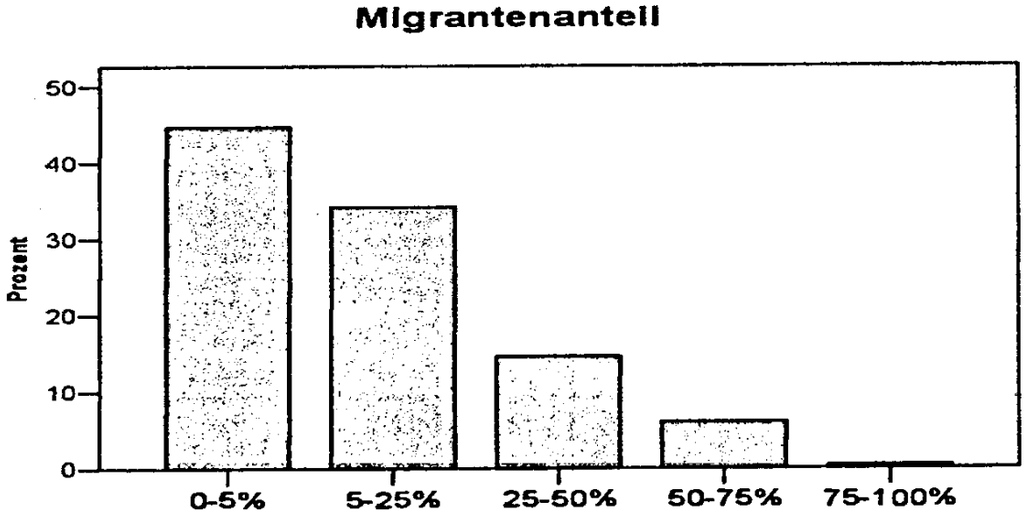
4.2 Schulische Kontextfaktoren

Für die Grundschulen wurden signifikante Effekte in Abhängigkeit vom Migrantenanteil und der Schulgröße gefunden. Regionale Unterschiede spielen bei den Grundschullehrkräften keine Rolle. Bei den Sekundarschullehrern gibt es lediglich signifikante Effekte in Abhängigkeit von der Schulgröße, nach dem Migrantenanteil wurde hier nicht gefragt.[5] Die beiden Kontextfaktoren Einzugsgebiet der Sekundarschule (ländlich vs. städtisch) und Ganztagesbetreuung haben dagegen keine Auswirkung auf die Akzeptanz und Einschätzung zentraler Lernstandserhebungen.

Einstellung zu Diagnosearbeiten in Abhängigkeit des Migrantenanteils

Der Migrantenanteil an Grundschulen streut sehr stark (vgl. Abbildung 3). Allerdings sind Schulen mit sehr hohem Migrantenanteil (über 50%) eher selten.

Abbildung 3: Migrantenanteil der Grundschulen, an denen die befragten Grundschullehrkräfte unterrichten.



Um Gruppen mit annähernd vergleichbaren Häufigkeiten zu erhalten, wurden die drei wenig besetzten Kategorien mit hohem Migrantenanteil zusammengefasst. Somit entstand eine Einteilung der Grundschullehrkräfte in drei vergleichbare Gruppen:

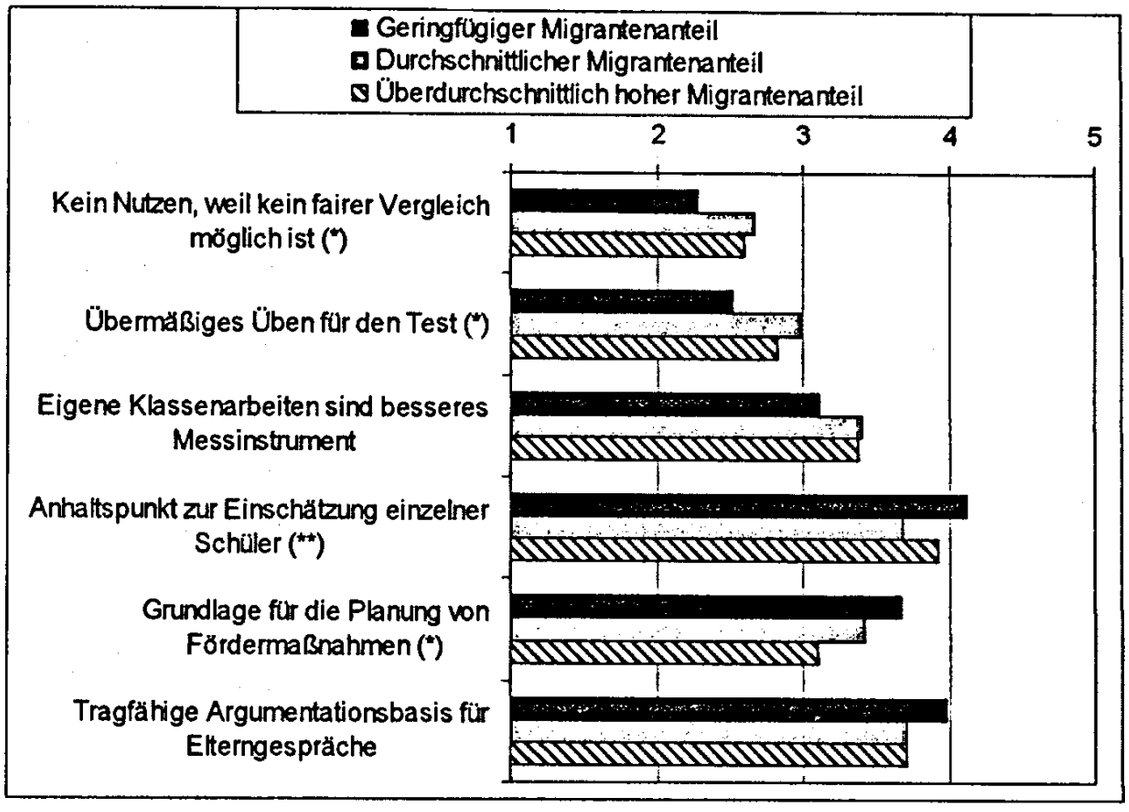
- Gruppe 1 (n=110): Lehrkräfte, die an einer Grundschule mit geringfügigem Migrantenanteil unterrichten (0-5%).
- Gruppe 2 (n=82): Lehrkräfte, die an einer Grundschule mit durchschnittlichem Migrantenanteil unterrichten (5-25%).
- Gruppe 3 (n=51): Lehrkräfte, die an einer Grundschule mit hohem Migrantenanteil unterrichten (über 25%).

---

5 In diesen Stichproben konnten diese Kontextdaten nicht erhoben werden.

Mit dieser Gruppeneinteilung wurde eine einfaktorielle Varianzanalyse über die Skala „allgemeine Akzeptanz zentraler Tests" und die Einzelitems zum förderdiagnostischen Nutzen und zu Risiken von Diagnosearbeiten gerechnet. Die allgemeine Akzeptanz zentraler Tests ist demnach bei Lehrkräften an Grundschulen mit geringfügigem Migrantenanteil (3,42) signifikant höher als bei Kolleginnen und Kollegen an Grundschulen mit durchschnittlichem (3,10) und hohem Migrantenanteil (3,23). Alle Werte bewegen sich jedoch weiterhin im positiven, zustimmenden Bereich.

Abbildung 4: Einfaktorielle Varianzanalyse über die Items zur Einschätzung der Diagnosearbeiten



Auch der mögliche Vorwurf unfairer Vergleiche sowie die Gefahr des übermäßigen Übens für den Test werden von den Grundschullehrkräften mit wenig Migrantenkindern deutlich stärker abgelehnt (Abbildung 4), während sie den förderdiagnostischen Nutzen der Diagnosearbeiten besonders hoch bewerten. Vor allem wenn es um die Planung von Fördermaßnahmen aufgrund der zentralen Lernstandsmessung geht, sind Lehrerinnen und Lehrer an Schulen mit einem überdurchschnittlich hohen Anteil an Migranten vergleichsweise skeptisch. Das deutet auf eine geringe Adaptivität der bislang eingesetzten Instrumente hin.

5. Diskussion und Ausblick

Durch die freiwillige Erprobung von Diagnose- und Vergleichsarbeiten in Baden-Württemberg in den Jahren 2003 und 2004 war ein Vergleich von Lehrereinstellungen zu zentraler Lernstandserhebungen in Abhängigkeit der bereits vorhandenen Testerfahrung möglich. Auch wenn diesen Diagnose- und Vergleichsarbeiten der Ernstcharakter fehlte und die Rückmeldeformate in ihrer

Komplexität keineswegs dem entsprachen, was mittlerweile bei Vergleichsarbeiten üblich ist, können die Ergebnisse der Befragung dennoch erste Hinweise geben, wie Lehrkräfte auf dieses neue Instrument reagieren und welchen Nutzen sie darin sehen.

Insgesamt äußerten die befragten Lehrerinnen und Lehrern eine eher positive Meinung gegenüber zentralen Lernstandserhebungen. Durch Testerfahrung wurde diese positive Grundeinstellung stabilisiert bzw. weiter ausgebaut. Mögliche Kritikpunkte werden nur von einer Minderheit der Lehrkräfte bestätigt. Vor allem Lehrkräfte mit Testerfahrung sehen deutlich weniger Risiken. Überwiegend positiv wird auch der förderdiagnostische Nutzen eingeschätzt, wenngleich die Ableitung von Fördermaßnahmen nicht so gut gelingt, wie vielleicht von Teilen der Lehrerschaft erwartet wurde. Es zeigte sich ebenfalls, dass die Einstellungen zu zentralen Tests von den Schulstufen unabhängig sind. Dieses ermutigende Zwischenfazit korrespondiert mit den bisherigen empirischen Befunden zum Umgang mit Leistungsrückmeldungen. Es bestätigt beispielsweise die Erwartung von Helmke (2004), dass schul- und klassenspezifische Rückmeldungen auf eine große Resonanz bei Lehrkräften stoßen müssten. Auch Groß Ophoff et al. (2006) interpretieren die Ergebnisse der VERA-Begleiterhebung in diese Richtung. Die hier berichteten Daten bestätigen somit ein insgesamt optimistisches Gesamtklima an Schulen bezüglich externer Leistungserhebungen. Die grundlegenden Einwände bezüglich einer schulinternen Nutzung extern erhobener Leistungsdaten (z.B. Rolff 2001) lassen sich mit diesen „klimatischen" Ergebnissen sicherlich kaum entkräften, denn wir haben schulische Prozesse und ihre Auswirkungen auf den Unterricht in dieser Studie nicht untersucht.

Weitere Ergebnisse der Studie weisen aber auch auf tiefer liegende Probleme bei der Einführung von Standards und zentralen Testuntersuchungen hin. Zunächst einmal ist eine stärkere Polarisierung bei Lehrkräften zu beobachten, die bereits mit Diagnose- und Vergleichsarbeiten Erfahrung sammeln konnten. Diese Spaltung innerhalb der Lehrerschaft bezüglich Testuntersuchungen wurde bereits von Ditton und Merz (2000) festgestellt und scheint sich durch die aktuellen Entwicklungen weiter zu verschärfen. Auch innerhalb von Lehrerkollegien an einer Schule werden vermutlich stark divergierende Meinungen aufeinander prallen. In der VERA-Begleiterhebung wird diese Problematik ebenfalls angedeutet. Groß Ophoff et al. (im Druck) resümieren, dass „die aus VERA abgeleiteten Maßnahmen nicht über die Klasse hinausgehen" und „dass die innerschulische Kooperation einen gewissen Wirkungskreis nicht überschreitet". Bezieht man diese Ergebnisse auf die eingangs thematisierten Wirkungsmodelle für Standards und Leistungsrückmeldungen, steht vielen Schulen womöglich noch ein langer Weg zur erhofften Evaluationskultur bevor. Sowohl O'Day (2004) als auch Visscher/Coe (2003) betonen in ihren Wirkungsmodellen die innerschulische Kommunikation über Leistungsdaten und die kooperative Ableitung gemeinsamer Handlungsziele.

Als relevante Kontextfaktoren für die Akzeptanz der Vergleichsarbeiten konnte vor allem der Migrantenanteil identifiziert werden. An Grundschulen mit einem zu vernachlässigenden Anteil an Migrantenkindern werden zentrale Lernstandserhebungen stärker befürwortet. An diesen Grundschulen sehen die Lehrkräfte auch einen größeren förderdiagnostische Nutzen von Diagnosearbeiten.

Die höhere Belastung der Lehrkräfte durch einen erhöhten Migrantenanteil könnte möglicherweise zu einer eher ablehnenden Haltung gegenüber Eingriffen bzw. Vorgaben „von oben" führen. Plausibel scheint auch die Hypothese, dass die bisher verfügbaren Diagnosearbeiten für durchschnittliche Grundschulklassen konzipiert wurden und nicht für die Diagnose sprachlicher Defizite und für die Ableitung entsprechender Fördermaßnahmen in Klassen mit hohem Migrantenanteil geeignet sind. Hierauf sollte die Testentwicklung durch Bereitstellung spezifischer Informationen zukünftig eingehen.

Auch der Durchführungszeitpunkt der baden-württembergischen Diagnose- und Vergleichsarbeiten trägt nicht unbedingt zu einer besseren schulinternen Verarbeitung der Ergebnisse bei. Die hier thematisierten Lernstandserhebungen wurden jeweils zum Ende der Schuljahre 2002/03 und 2003/04 durchgeführt und haben dadurch den Charakter einer summativen Evaluation. Die in den Klassen 3 und 5 durchgeführten Diagnose- und Vergleichsarbeiten eröffnen noch am ehesten die Möglichkeit, Konsequenzen für den weiteren Lern- und Unterrichtsverlauf abzuleiten, da hier in der Regel kein Lehrerwechsel stattfindet. Zentrale Lernstandserhebungen, die der Schul- und Unterrichtsentwicklung dienen sollen, müssten dagegen während eines laufenden Schuljahres durchgeführt werden, um eine zeitnahe Ableitung geeigneter Maßnahmen zu ermöglichen (v. Ackeren und Bellenberg 2004).

Die Schuladministration könnte durch den Aufbau geeigneter und informativer Rückmeldesystem, z.B. in der Art des Schweizer Klassencockpits, die insgesamt günstige Grundstimmung der Lehrkräfte aufgreifen und damit eine neue Qualität in den Prozess der Schulentwicklung bringen. Die Wirkungen solcher Systeme zu untersuchen, wäre dann eine Aufgabe der Evaluationsforschung. Um das komplexe Zusammenspiel zwischen der individuellen Bereitschaft einer einzelnen Lehrkraft und einer kollektiv geteilten Verantwortlichkeit gegenüber Leistungsrückmeldungen an Schulen zu untersuchen, sollten Rezeptionsstudien vor allem durch qualitative Studien auf der schulischen Mikroebene ergänzt werden. Diese hätten sich vor allem mit dem Zusammenhang zwischen externer Datenrückmeldung und interner Evaluationskultur zu beschäftigen (vgl. Rolff 2001) und könnte dann Gelingensbedingungen an einzelnen Schulen herausarbeiten.

Wir danken dem Forschungsverbund Hauptschule sowie der Landesstiftung Baden-Württemberg für die Finanzierung der hier vorgestellten Studien.

Literatur

Abelmann, Charles; Richard Elmore 1999: When accountability knocks, will anyone answer? (CPRE Research Report No. RR-42). Philadelphia: Consortium for Policy Research in Education, University of Pennsylvania

van Ackeren, Isabell; Gabriele Bellenberg 2004: Parallelarbeiten, Vergleichsarbeiten und Zentrale Abschlussprüfungen. In: Heinz-Günter Holtappels u.a. (Hrsg.): Jahrbuch der Schulentwicklung, Band 13, IfS. Weinheim: Juventa, S. 125-159

Arnold, Karl-Heinz 2002: Schulentwicklung durch Rückmeldung der Lernwirksamkeit an die Einzelschule: Möglichkeiten und Grenzen der Schuleffizienzforschung. In: Zeitschrift für Pädagogik, 48, 2002, 5, S. 741-764

Bonsen, Martin; Jan von der Gathen 2004: Schulentwicklung und Testdaten. Die innerschulische Verarbeitung von Leistungsrückmeldungen. In: Hans-Günter Rolff u.a. (Hg.): Jahrbuch der Schulentwicklung, Band 13, IfS, Weinheim: Juventa, S. 225-252

Böttcher, Wolfgang 2002: Kann eine ökonomische Schule auch eine pädagogische sein? Schulentwicklung zwischen Neuer Steuerung, Organisation, Leistungsevaluation und Bildung. Weinheim und München: Juventa
Böttcher, Wolfgang 2004: Bildungsstandards und Kerncurricula – Potenzielle, intendierte und nicht-intendierte Effekte eines zentralen Reformprojektes. In: Die Deutsche Schule, 8. Beiheft, 2004, S.231-244
Coe, Robert 2002: Evidence on the role and impact of performance feedback in schools. In: Adrie J. Visscher, Robert Coe (Eds.): School improvement through performance feedback. Lisse: Swets & Zeitlinger, S. 3-26
DeBray, Elizabeth, Gail Parson, Katrina Woodworth 2001: Patterns of response in four high schools under state accountability policies in Vermont and New York. In: Susan H. Fuhrman (Ed.): From the capitol to the classroom: Standards-based reform in the states. Chicago: University of Chicago Press, pp. 170-192
Ditton, Hartmut; Daniela Merz 2000: Qualität von Schule und Unterricht – Kurzbericht über erste Ergebnisse einer Untersuchung an bayerischen Schulen. Katholische Universität Eichstätt / Universität Osnabrück
Ditton, Hartmut 2002: Evaluation und Qualitätssicherung. In: Rudolf Tippelt (Hg.): Handbuch Bildungsforschung. Opladen: Leske + Budrich, S. 775-790
Groß Ophoff, Jana; Ursula Koch, Ingmar Hosenfeld, Andreas Helmke 2006: Ergebnisrückmeldung und ihre Rezeption im Projekt VERA. In: Harm Kuper, Julia Schneewind (Hg.): Rückmeldung und Rezeption von Forschungsergebnissen. New York, München, Berlin: Waxmann, S. 19-40
Groß Ophoff, Jana; Ursual Koch, Ingmar Hosenfeld, Andreas Helmke (im Druck): Das Projekt VERA: Von der Evaluation zur Schul- und Unterrichtsentwicklung! In: Schul-Verwaltung HE/RP, 2006, Heft 5 und 6
Helmke, Andreas; Ingmar Hosenfeld 2005: Standardbezogene Unterrichtsevaluation. In: Gerold Brägger, Beat Bucher, Norberg Landwehr (Hg.): Schlüsselfragen zur externen Schulevaluation. Bern: Hep Verlag, S. 127-151
Helmke, Andreas 2003: Unterrichtsqualität: erfassen – bewerten – verbessern. Seelze: Kallmeyer
Helmke, Andreas 2004: Von der Evaluation zur Innovation. Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. In: Seminar, 2004, 2, S. 90-112
Klieme, Eckhard 2004: Begründung, Implementation und Wirkung von Bildungsstandards: Aktuelle Diskussionslinien und empirische Befunde. In: Zeitschrift für Pädagogik, 50, 2004, 5, S. 625-634
Klieme, Eckhard; u.a. 2003: Zur Entwicklung nationaler Bildungsstandards – Eine Expertise. Berlin: Bundesministerium für Bildung und Forschung
Künzli, Rudolf 1999: Lehrplanarbeit – Steuerung von Schule und Unterricht. In: Rudolf Künzli, u.a. (Hg.): Lehrplanarbeit – Über den Nutzen von Lehrplänen für die Schule und ihre Entwicklung. Zürich: Ruegger, S. 11-30
Nachtigall, Christof & Jantowski, Andreas 2004: Die Thüringer Kompetenztests. In: Neue Praxis der Schulleitung, Thüringen, 2004, 73, S. 1-14
Nachtigall, Christof; Ulf Kröhne 2006: Methodische Anforderungen an schulische Leistungsmessung – auf dem Weg zu fairen Vergleichen. In: Harm Kuper, Julia Schneewind (Hg.): Rückmeldung und Rezeption von Forschungsergebnissen. New York, München, Berlin: Waxmann; S. 59-74
O'Day, Jennifer A. 2002: Complexity, accountability, and school improvement. In: Harvard Educational Review, 72, 2002, 3, pp. 293-329
O'Day, Jennifer A. 2004: Complexity, accountability, and school improvement. In: Susan H. Fuhrman, Richard Elmore (Eds.): Redesigning Accountability Systems for Education. New York, London: Teachers College Press, pp. 15-43
Oelkers, Jürgen 2003: Wie man Schule entwickelt – Eine bildungspolitische Analyse nach PISA. Weinheim: Beltz
Peek, Rainer; Peter Dobbelstein 2006: Benchmarks als Input für die Schulentwicklung – das Beispiel der Lerstandserhebungen in Nordrhein-Westfalen. In: Harm Kuper,

Julia Schneewind (Hg.): Rückmeldung und Rezeption von Forschungsergebnissen. New York, München, Berlin: Waxmann, S. 41-58
Rauin, Udo; Klaus-Jürgen Tillmann, Witloff Vollstädt 1996: Lehrpläne, Schulalltag und Schulentwicklung. In: Hans-Günter Rolff, u.a. (Hg.): Jahrbuch der Schulentwicklung, Band 9, Weinheim, München: Juventa, S. 377-414
Rolff, Hans-Günter 2001: Was bringt die vergleichende Leistungsmessung für die pädagogische Arbeit an Schulen? In: Franz E. Weinert (Hg.): Leistungsmessungen in Schulen. Weinheim, Basel: Beltz, S. 337-365
Rossi, Peter H.; Howard E. Freeman 1993: Evaluation – A systematic approach. Newbury Park, London, New Dehli: Sage
Scheerens, Jaap, Cees Glas, Sally M. Thomas 2003: Educational evaluation, assessment, and monitoring – a systemic approach. Lisse: Swets & Zeitlinger
Schrader, Friedrich-Wilhlem, Andreas Helmke 2004: Von der Evaluation zur Innovation? Die Rezeptionsstudie WALZER: Ergebnisse der Lehrerbefragung. In: Empirische Pädagogik, 18, 2004, 1, S. 140-161
Terhart, Ewald 2002: Wie können die Ergebnisse von vergleichenden Leistungsstudien systematisch zur Qualitätsverbesserung in Schulen genutzt werden? In: Zeitschrift für Pädagogik, 48, 2002, 1, S. 91-110
Vollstädt, Witloff, Klaus-Jürgen Tillmann, Udo Rauin, Katrin Höhmann, Andrea Tebrügge 1999: Lehrpläne und Schulalltag. Eine empirische Studie zur Akzeptanz und Wirkung von Lehrplänen in der Sekundarstufe I. Opladen: Leske + Budrich

Uwe Maier, geb. 1971, Dr., Akademischer Rat; Schwerpunkte: Lehrpläne, Bildungsstandards, Übergänge;
Anschrift: Pädagogische Hochschule Schwäbisch Gmünd; Oberbettringerstraße 200; 73525 Schwäbisch Gmünd;
Email: uwe.maier@ph-gmuend.de;
Website: http://schulpaedagogik.ph-gmuend.de

Udo Rauin geb. 1954, Dr., Professor; Schwerpunkte: Schul- und Unterrichtsforschung, Lehrplanforschung, Evaluationsforschung; Institut für Pädagogik der Sekundarstufe an der Johann Wolfgang Goethe-Universität Frankfurt a.M.;
Anschrift: Universität Frankfurt, Fachbereich Erziehungswissenschaften, Institut für Pädagogik der Sekundarstufe, Postfach 111932, 60054 Frankfurt a.M.;
Email: rauin@em.uni-frankfurt.de;
Website: http://www.uni-frankfurt.de/fb/fb04/personen/rauin.html