Maier, Uwe

Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften

Zeitschrift für Pädagogik 54 (2008) 1, S. 95-117

ZEITSCHRIFT FÜR PÄDAGOGIK
Jahrgang 54 – Heft 1
Januar/Februar 2008

Inhaltsverzeichnis

Allgemeiner Teil

Besprechungen

Dokumentation

Uwe Maier

Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften

Zusammenfassung: *Vergleichsarbeiten sollen die datenbasierte Schul- und Unterrichtsentwicklung an Schulen initiieren und stützen. Voraussetzung hierfür ist die Akzeptanz zentraler Tests durch Lehrkräfte und die gezielte Nutzung der Leistungsrückmeldungen für bestimmte Bereiche des unterrichtlichen Handelns. In einer Studie direkt im Anschluss an die erste verpflichtende Vergleichsarbeitsrunde in Baden-Württemberg wurden (n=307) Lehrkräfte zum pädagogischen Nutzen der rückgemeldeten Leistungsdaten sowie zu möglichen Kontextfaktoren schriftlich befragt. Die Ergebnisse weisen vor allem auf schulformspezifische Differenzen in der Wahrnehmung von Vergleichsarbeiten hin. Lehrkräfte sehen überdies eher eine selektionsdiagnostische Nutzung der Leistungsdaten, obwohl die Akzeptanz der Vergleichsarbeiten stark mit der förderdiagnostischen Nutzung zusammenhängt.*

1. Einleitung

Die Einführung zentraler Vergleichsarbeiten konfrontiert Schulen mit einem Kernelement der neuen, ergebnisorientierten Schulsystemsteuerung. Mit landesweiten, standardisierten Tests soll das Leistungsniveau der Schüler an Bildungsstandards als zentraler Norm gemessen werden. Aus bildungspolitischer Sicht steht dabei die Umsetzung des ergebnisorientierten Steuerungsparadigma im Vordergrund. Aber die zentrale Erfassung von Schulleistungen führt noch lange nicht zu einem verbesserten Unterricht und somit zu besseren Schülerleistungen. Völlig ungeklärt sind beispielsweise die Fragen, welche Bereiche des unterrichtlichen Handelns durch Leistungsrückmeldung an Lehrkräfte überhaupt beeinflusst werden können und von welchen Kontextfaktoren die Akzeptanz und somit auch Nutzbarmachung zentraler Vergleichsarbeiten abhängt. Die vorliegende Studie möchte zur Klärung dieser Fragen einen Beitrag leisten. In der theoretischen Hinführung wird zunächst ausgelotet, unter welchen Umständen eine Anbindung zentraler Leistungsvergleichsstudien an Schulentwicklungsprozesse möglich und sinnvoll ist. Anschließend werden empirische Forschungsergebnisse zur Rezeption und Nutzung externer Leistungsrückmeldungen aus dem deutschsprachigen Raum dargestellt und Forschungsdesiderata herausgearbeitet.

1.1 Vom Systemmonitoring zur Einzelschulentwicklung

Nationale und internationale Leistungsvergleichsstudien haben in Deutschland den gesellschaftlichen Diskurs über Bildungswirkungen in Schwung gebracht. Das inzwischen etablierte Systemmonitoring verspricht verlässliche Daten für die Bildungsplanung zur Verfügung zu stellen. Diese Form der zentralen Evaluation hat aber keine direkt steu-

ernde Funktion für Einzelschulen (Scheerens/Glas/Thomas 2003). Rückwirkungen auf Schule und Unterricht sind nur über Vermittlungsinstitutionen mit administrativer Legitimation wie z.B. Lehrpläne oder die Lehrerbildung denkbar (Baumert 2001; Klieme u.a. 2003). Rolff (2001) unterscheidet in diesem Zusammenhang auch zwischen vermittelndem und unmittelbarem Nutzen. Stichprobenuntersuchungen wie PISA oder IGLU können für Lehrkräfte und Schulen allenfalls einen vermittelnden Nutzen haben, weil sie keine spezifische Information für Einzelschulen bereitstellen. Diese Studien produzieren Orientierungswissen und liefern Impulse für die fachdidaktische Diskussion. Ein unmittelbarer Nutzen für Schulen und Lehrkräfte kann allenfalls von flächendeckenden Lernstandserhebungen erwartet werden.

Die flächendeckende Evaluation von Einzelschulen oder ausgewählten Jahrgängen orientiert sich methodologisch an der internationalen Schulleistungsforschung. Durch die inhaltliche Anbindung der Tests an die gültigen curricularen Vorgaben und die flächendeckende Durchführung wird jedoch auch die schulpraktische Relevanz der Leistungsrückmeldungen gewährleistet. Unter Berücksichtigung internationaler Erfahrungen unterscheidet Baumert (2001) zwischen zwei Formen der flächendeckenden Evaluation. Vor allem im angloamerikanischen Bereich dominiert das Wettbewerbsmodell mit der Grundannahme einer nachfragegesteuerten Qualitätsentwicklung, ausgelöst durch eine Veröffentlichung von schulischen Performanzindikatoren. Nach Baumert hat dieses Modell in den deutschsprachigen europäischen Ländern keine Anhänger. Die zweite Variante ist das Modell professioneller Qualitätsentwicklung. Die Leistungsinformationen werden nicht veröffentlicht, sondern dienen den Einzelschulen als Datengrundlage für den professionellen Diskurs über die Lernentwicklung der Schüler und geben Impulse für die systematische Weiterentwicklung der eigenen Arbeit.

Die in Deutschland eingeführten Vergleichsarbeiten entsprechen weitgehend dem letztgenannten Modell (Peek/Dobbelstein 2006a). Sie knüpfen somit an bereits vorhandene Formen der dezentralen Evaluation im Rahmen von Schulentwicklungsprozessen an. Aus schultheoretischer Sicht erzeugt gerade dies ein konfliktträchtiges Spannungsfeld. Einerseits wird mit zentralen Tests der Versuch unternommen, Schuleffektivitätsforschung für Einzelschulentwicklungsprozesse anschlussfähig zu machen. Andererseits werden die Tests von der Bildungsadministration verordnet und somit von Lehrkräften auch als externes Kontrollinstrument wahrgenommen. Die Nutzung für die Schulentwicklung setzt die Problembearbeitung und die Problemlösung von der Einzelschule voraus. Die Praktiker vor Ort sollen ihre Lage analysieren, sich selbst Reformziele setzen und geeignete Maßnahmen ergreifen. In diesem Prozess können zentrale Leistungsrückmeldungen ein wertvolles Instrument sein (z.B. Visscher/Coe 2003).

Aber spätestens dann, wenn so genannte „Produktdaten" über Unterricht erhoben werden, wird jede schulinterne Evaluation zu einem konfliktreichen Unterfangen, weil es Lehrkräfte zumindest implizit individuell evaluiert (Baumert 2001).

Hinzu kommt, dass das Verhältnis von externer und interner Evaluation auch ohne zentrale Lernstandserhebungen in den meisten Bundesländern noch ungeklärt ist (Rolff 2001; Böttcher/Holtappels/Brohm 2006). Wenn an Schulen bisher überhaupt evaluiert wurde, war diese interne Evaluation in der Regel nur selten mit externen Formen der

Evaluation gekoppelt. Im schlechtesten Fall kam es zum „Gegentakt" zwischen interner und externer Evaluation, wobei die Schule Außenkriterien für die Evaluation dezidiert ablehnt. Vergleichsarbeiten und die entsprechenden Leistungsrückmeldungen sind jedoch sehr strikte Außenkriterien, die nun aber zur Selbstevaluation genutzt werden sollen. Die Nutzung von Leistungsrückmeldungen in den einzelnen Lehrerkollegien ist somit sehr fragwürdig und es ist mit hochkomplexen Rezeptionsproblemen zu rechnen.

1.2 Nutzung zentraler Leistungsrückmeldungen auf Schul- und Unterrichtsebene

Die Verwendung von Rückmeldedaten für die Weiterentwicklung von Schule und Unterricht ist bildungspolitisch gewollt und lässt sich an das aktuelle Paradigma der Schulentwicklungsforschung koppeln. Die pädagogische Nutzung rückgemeldeter Schülerleistungen ist jedoch weitaus problematischer als von administrativer Seite angenommen wird. Weiss (1998) betont, dass Evaluationsinformationen partiell, fragmentarisch, mit Unterbrechungen, unangemessen oder überhaupt nicht genutzt werden können. Es stellt sich somit die Frage nach Modellen, die zu einer pädagogisch sinnvollen, angemessenen und kontinuierlichen Verwendung zentraler Leistungsrückmeldungen führen.

Ein strukturelles Rahmenmodell für die pädagogische Nutzbarmachung von Vergleichsarbeiten wurde von Helmke und Hosenfeld (2005) vorgelegt. Im Kern dieses Modells wird der Evaluationszyklus nachgebildet, wobei die Rückmeldeinformationen als Ausgangsbasis oder Impulsgeber angesehen werden. Der gesamte Zyklus von der Rezeption und Interpretation der Leistungsdaten über die daraus resultierenden Aktionen bis hin zur erneuten Evaluation der Handlungen wird von individuellen und schulischen Rahmenbedingungen moderiert. Das Modell eignet sich, um Phasen der Nutzbarmachung und externe Einflussfaktoren zu lokalisieren. Für eine empirische Überprüfung fehlt jedoch die inhaltliche Spezifikation von Zusammenhängen.

Eine einfache Typisierung der Nutzung von Leistungsdaten geht auf Rossi und Freeman (1993) zurück. Sie unterscheiden zwischen direkter, konzeptueller und symbolischer Nutzung von Feedbackinformationen. Günstig ist die direkte oder auch instrumentelle Nutzung, bei der aufgrund der zur Verfügung stehenden Leistungsdaten konkrete Handlungsentscheidungen getroffen werden. Eine konzeptuelle Nutzung liegt dann vor, wenn die Rückmeldung bewertet wird und damit das Denken der Entscheidungsträger indirekt beeinflusst. Wird die Feedbackinformation lediglich selektiv genutzt, um den eigenen, bereits feststehenden Standpunkt argumentativ zu stützen, sprechen Rossi und Freeman von einer symbolischen Nutzung. Diese Typisierung ermöglicht eine erste grobe Einschätzung von Umgangsweisen mit zentralen Leistungsrückmeldungen. Unbeantwortet bleibt allerdings die Frage nach der Gestaltung von Leistungsrückmeldungen, die eine instrumentelle Nutzung bedingen können.

Hierauf geben Kluger und DeNisi (1996) mit ihrer allgemeinen, nicht nur auf die Nutzung zentraler Leistungsrückmeldungen bezogenen *feedback intervention theory* eine abstrakte Antwort. Danach beeinflussen vor allem drei Variablenklassen die Wirksamkeit einer Rückmeldung: Zusätzliche Hinweise in der Rückmeldung, das Wesen der

gemessenen Leistung und situative bzw. personale Variablen. Beispielsweise können Lob und Tadel als zusätzliche Hinweise in der Rückmeldung die Aufmerksamkeit von der eigentlichen Information ablenken. Rückmeldungen sollten eine internale Attribuierung der Leistungsergebnisse unterstützen, um Handlungsmotivation zu erzeugen. Bezüglich der gemessenen Leistung, auf die sich die Rückmeldung bezieht, stellen Kluger und DeNisi die Aufgabenkomplexität als Hauptfaktor heraus. Für einfache Aufgaben haben Rückmeldungen in der Regel substanzielle Effekte. Bei komplexen Aufgaben findet man eher keine Effekte, bzw. kann sie nur schlecht nachweisen.

Wesentlich näher am Gegenstandsbereich sind die systemtheoretisch orientierten Überlegungen von O'Day (2002; 2004) zur Wirksamkeit von so genannten *school accountability systems*[1] in den USA. Ein komplexes, soziales System lernt dann, wenn durch die Verarbeitung von Information die Bandbreite der Handlungsmöglichkeiten erweitert wird. Für das System Schule bedeutet dies, dass Lehrer lernen, relevante Informationen für zukünftiges Handeln von ineffektiv erscheinender zu unterscheiden. Des Weiteren müssen Informationen im System richtig interpretiert werden, um Entwicklungen in Gang setzen zu können. Dies ist in komplexen Systemen wie der Schule nicht immer der Fall. Einschränkungen ergeben sich z.B. durch die individuell unterschiedliche Wissensbasis der Mitglieder, sozial konstruierte Überzeugungen (*belief systems*) und die Machthierarchie innerhalb der Institution.

1.3 Empirische Forschungsergebnisse zur Akzeptanz und Nutzung zentraler Leistungsrückmeldungen in Schulen

Parallel zur öffentlichen TIMSS- und PISA-Diskussion wurden erste Lehrerbefragungen zur Rezeption großer Leistungsvergleichsstudien durchgeführt (z.B. Kohler 2004). Dabei zeigte sich durchweg, dass die Rückmeldung hoch aggregierter Daten aus Stichprobenstudien von Lehrkräften – im Sinne von Rossi und Freeman (1993) – allenfalls symbolisch genutzt wird. Etwas optimistischer sind die Ergebnisse einer neueren Studie zur PISA-Akzeptanz aus Sachsen (Sedlmeier/Böhm/Lindner/Schmidt 2006), wo es nicht nur zu einseitigen externalen Attribuierungen kam und wo die von den Lehrkräften geäußerten Verbesserungsvorschläge gut mit den Befunden der Lehr-Lernforschung korrespondierten. Auch Imhof (2005) befragte Lehrer in verschiedenen schulischen Funktionen zu den PISA-Ergebnissen. Die Informiertheit der Probanden variierte sehr stark und die Informationsangebote wurden sehr unterschiedlich genutzt. Es gab allerdings einen engen Zusammenhang zwischen Informiertheit und der Bereitschaft, sich an Maßnahmen zur Qualitätsverbesserung zu beteiligen. Aufgrund der nicht intendierten

1 *School accountability systems* sind die pädagogischen Antwort auf rein marktorientierte Systeme der Rechenschaftslegung und zielen auf eine schulinterne Evaluation, gestützt durch externe, landesweite Leistungstests. Damit sind diese Systeme mit dem vergleichbar, was im deutschsprachigen Raum momentan unter den Stichworten „Vergleichsarbeiten" und „Schulevaluation" diskutiert wird.

Kopplung von Leistungsrückmeldung und Schulentwicklung können diese ersten Studien zur Wahrnehmung und Akzeptanz internationaler Schulleistungsvergleiche nur eingeschränkte Befunde liefern. Zudem sind die Zeitabstände zwischen Leistungsmessung, Datenrückmeldung und Lehrerbefragung zur Nutzung der Daten in der Regel recht groß.

Diese Defizite treten bei „Rezeptionsstudien" im Zuge nationaler Leistungsvergleiche mit Gesamterhebungen nicht mehr auf. Allerdings unterscheiden sich diese Studien in Design, Stichprobenumfang und methodischem Instrumentarium zum Teil erheblich, sodass direkte Vergleiche und Generalisierungen nur schwer möglich sind. Peek (2004) konnte beispielsweise zeigen, dass die Rezeption der QuaSUM[2]-Ergebnisse an den Schulen vor allem fachbezogen erfolgt. In den Fachkonferenzen der Mathematiklehrkräfte fand eine intensive Auseinandersetzung mit den Ergebnissen der Leistungsstudie statt. Dieser Befund deckt sich weitgehend mit den Bedingungen, unter denen Schulentwicklungsprozesse gelingen (Rolff 2001). Die Fachkollegien haben die Kompetenz, um wirksame Konsequenzen aus zentralen Leistungsdaten zu ziehen. Datenbasierte Unterrichtsentwicklung hängt damit natürlich auch von wirksamen kollegialen Kooperationsformen innerhalb einer Schule ab. Indirekt wird dieses Ergebnis durch eine qualitative Studie über die innerschulische Rezeption von extern erhobenen Leistungsdaten der Hamburger LAU-Studie[3] bestätigt (Klug/Reh 2000). Dort sahen die – von den Fachkonferenzen weiter entfernten – Schulleiter kein Erklärungspotenzial der zurückgemeldeten Leistungsdaten für eine schulinterne Ursachenforschung.

Wesentlich ernüchternder sind dagegen die Ergebnisse der an MARKUS angegliederten WALZER-Befragung[4] (Schrader/Helmke 2004). Die überwiegende Mehrheit der Lehrkräfte hat keine Konsequenzen aus der Datenrückmeldung gezogen. Die Autoren plädieren deshalb dafür, die Erwartungen nach unten zu korrigieren und fordern weitere Maßnahmen, um eine evaluationsbasierte Schulentwicklung zum Erfolg zu führen. Auch bei VERA[5] wurden die beteiligten Grundschullehrkräfte zum Umgang mit Vergleichsarbeiten und den Auswirkungen auf Unterrichtsentwicklungsmaßnahmen befragt (Groß Ophoff/Koch/Hosenfeld/Helmke 2006). Die Tests wurden von den Lehrkräften als informativ wahrgenommen und die Durchführung sowie die Auswertung bereiteten kaum Probleme. Im Vordergrund stand aus Lehrersicht die Ableitung von geeigneten Fördermaßnahmen für einzelne Schüler. Die aus VERA abgeleiteten Maßnahmen gingen allerdings nicht über die Klasse hinaus oder beschränkten sich lediglich auf bereits praktizierte Formen kollegialer Kooperation.

Moser (2003) befragte 216 Lehrerinnen und Lehrer der 6. Klassen im Kanton Zürich über erste Erfahrungen mit dem „Klassencockpit". Das Klassencockpit wird von einem

2 Qualitätsuntersuchungen an Schulen zum Unterricht in Mathematik (Brandenburg)
3 Untersuchung zur Lernausgangslage an Hamburger Schulen
4 Wirkungsanalyse der Leistungsevaluation: Zielerreichung, Ertrag für die Bildungsqualität der Schule und die Rückmeldung von Evaluationsergebnissen
5 Vergleichsarbeiten in der Grundschule

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. Zwei Drittel der Lehrkräfte erklärten, an einer Optimierung des eigenen Unterrichts mithilfe externer Leistungsvergleiche interessiert zu sein. Fast alle (95%) waren an einem sozialen Vergleich mit den kantonalen Mittelwerten interessiert. Auch die Ergebnisse zur tatsächlichen Nutzung der Vergleichsdaten weisen auf Prioritäten hin. 89% der Befragten nutzten die Daten als Bestätigung der eigenen Schülerbeurteilungen, 82% als Ergänzung der eigenen Schülerbeurteilung und jede fünfte Lehrkraft gab sogar an, die eigenen Noten aufgrund der Testergebnisse revidiert zu haben. Die Nutzung der Daten für Diskussionen im Kollegium lag bei 45% und jede dritte Lehrkraft kreuzte an, aufgrund der Leistungsrückmeldung über den eigenen Unterricht reflektiert zu haben.

Rückmeldeinformationen scheinen zunächst einmal die Reflexion über die eigene Leistungsbeurteilung anzuregen. Die Reflexion über Unterricht auf der Grundlage von Schülerleistungsdaten erfordert hingegen weitere interpretative Zwischenschritte und zusätzliche Informationen. Eine ausführliche Aufgaben- bzw. Fehleranalyse vor dem Hintergrund der jeweiligen Schülerlernvoraussetzungen ist ein entscheidendes Bindeglied zwischen Leistungsrückmeldungen und Reflexion des Unterrichts. Trotz zentraler Vorschläge in diesem Bereich muss dies die Lehrkraft im Wesentlichen selbst leisten. Eine weitere Studie aus der Schweiz liefert ebenfalls Belege für die These, dass sich zentrale Tests am ehesten auf die selektionsdiagnostische Praxis auswirken können. Werden standardisierte Tests als zusätzliche Entscheidungsgrundlage für die Bildungsempfehlung am Ende der Primarstufe eingesetzt, so geht der Einfluss des familialen Hintergrunds auf die Übergangsentscheidung zurück (Baeriswyl/Wandeler/Trautwein/Oswald 2006).

2. Forschungsdesiderata und Fragestellungen

Ein zentrales Problem aller bisher berichteten Studien zur Rezeption von Leistungsrückmeldungen sind die unterschiedlichen Zeitabstände zwischen Test, Datenrückmeldung und Befragung zur Datenrückmeldung. Aussagen der Lehrkräfte zur Nutzung von Leistungsinformationen werden aber vermutlich vom Faktor Zeit nicht unabhängig sein. Des Weiteren unterscheidet sich die Form der Leistungsmessung bzw. Datenrückmeldung (zentral vs. dezentral, usw.) ebenfalls zum Teil erheblich. Als weitere Kritik an der bisherigen Forschlage kommt hinzu, dass die stichprobenmäßig großen Studien in der Regel von Forschergruppen durchgeführt wurden, die auch für die zentrale Leistungsmessung verantwortlich waren.

Unabhängig von diesen methodischen Einschränkungen legen die bisherigen empirischen Studien als auch die theoretischen Überlegungen eine grundlegende These nahe: Lehrkräfte können externe Leistungsrückmeldungen interpretieren und mit ihnen weiterarbeiten, allerdings nicht unbedingt im Sinne der externen Zielsetzungen, son-

dern immer nur vor dem Hintergrund nahe liegender Handlungserfordernisse der täglichen Unterrichtspraxis. Die Bildungsadministrationen tragen dieser These unbewusst Rechnung, indem sie die Nutzung der Rückmeldeinformationen nicht detailliert vorschreiben. Die bisherige Forschungslage reicht jedoch nicht aus, um eindeutig beschreiben zu können, welche schulischen Handlungsfelder und welche schulischen Akteure unter welchen Bedingungen mit externen Leistungsdaten produktiv arbeiten können. Dabei fehlen vor allem tragfähige Evaluationsstudien, die Auswirkungen von Lernstandserhebungen auf Schul- und Unterrichtsentwicklungsprozesse in den beteiligten Schulen untersuchen (Klieme 2004; Peek/Dobbelstein 2006b).

Beispielsweise stellt sich die Frage nach schulischen Handlungsfeldern, die neben der Selektionsdiagnostik für externe Leistungsdaten „sensibel" sind. Lässt sich diese differenzielle Sensibilität, die für interne Evaluation belegt ist, für externe Daten bestätigen? Ebenso ungeklärt ist, von welchen schulischen Kontextfaktoren die Nutzbarmachung von Leistungsrückmeldungen für die interne Schulentwicklung abhängt. In der Regel wird ein Typus von Vergleichsarbeiten für ein ganzes Bundesland entwickelt und eingesetzt. Interne Evaluation erfordert aber immer ein auf die Schule abgestimmtes Instrumentarium. Wissen über relevante Kontextfaktoren könnte somit als Grundlage für eine sinnvolle Diversifikation dienen. Dasselbe gilt für individuelle Kontextfaktoren. Nicht jede Lehrkraft wird in gleichem Maße über Kompetenzen und Einstellungen verfügen, um externe Leistungsdaten sinnvoll interpretieren zu können.

Aufgrund der Einführung verpflichtender Vergleichsarbeiten ab dem Schuljahr 2005/06 stellen sich diese Fragen auch für das baden-württembergische Schulsystem. Bereits ab 1999 gab es erste, vom Landesinstitut für Schulentwicklung durchgeführte Probeläufe für zentrale Leistungstests, an denen Lehrkräfte freiwillig teilnehmen konnten. Die flächendeckende Einführung der Vergleichsarbeiten ergänzte dann die im Zuge der Bildungsplanreform 2004 formulierten schulformspezifischen Bildungsstandards. Aus Sicht der baden-württembergischen Bildungsadministration sollen Vergleichsarbeiten den Lernstand der Schüler objektiv dokumentieren, diagnostische Informationen liefern aber auch Teil eines schulinternen und externen Qualitätssicherungssystems sein[6]. In einer gesonderten Handreichung werden den Schulen konkrete Vorschläge gemacht, welche Konsequenzen aus der Ergebnisinterpretation der Vergleichsarbeiten gezogen werden können[7]. Dabei werden folgende Handlungsfelder angesprochen: Diagnose und Förderung (z.B. Welche Schülerinnen und Schüler bedürfen in welchen Bereichen zusätzliche Hilfen?), Unterrichtsqualität (z.B. Welche Aspekte des Unterrichts bedürfen keiner Veränderung, welche sollten überdacht werden?) sowie Kooperations- und Teamstrukturen (In welcher Weise können kommunikative Strukturen gestärkt werden, um Diagnose und Förderung sowie Unterrichtsqualität weiterzuentwickeln?)

Vergleichsarbeiten in Baden-Württemberg stehen somit in dem bereits erläuterten Spannungsfeld zwischen externer Kontrolle und Selbstevaluation bzw. Schulentwick-

6 http://lbsneu.schule-bw.de/unterricht/dva/ [abgerufen am 11.1.2007]

7 http://lbsneu.schule-bw.de/unterricht/dva/dva_docs/umgang.pdf [abgerufen am 11.1.2007]

lung. Folgende Forschungsfragen sind deshalb Grundlage und Ausgangspunkt für die hier berichtete explorative Studie:

1) Wie schätzen Lehrkräfte die Nutzung zentraler Lernstandserhebungen für die zukünftige Unterrichtsgestaltung und die Unterstützung von selektions- und förderdiagnostischen Entscheidungen ein?
2) In welchem Maße hängt die Einschätzung der Nutzung zentraler Leistungsrückmeldungen von der Akzeptanz und der subjektiv wahrgenommenen Lehrplanvalidität der Tests ab?
3) Wie groß sind schulform- bzw. fachspezifische Differenzen in der Einschätzung der Akzeptanz und der Nutzung zentraler Lernstandserhebungen?
4) Welche individuellen, klassen-, schul- und regionalspezifischen Kontextvariablen beeinflussen die Sichtweise der Lehrkräfte auf die Nutzung zentraler Lernstandserhebungen?

3. Methodisches Vorgehen

Um die Einstellungen und Nutzungsmöglichkeiten aus Lehrerperspektive zu erfassen, wurde direkt im Anschluss an die erste verpflichtende Vergleichsarbeitsrunde in Baden-Württemberg eine schriftliche Befragung von Lehrkräften durchgeführt. In den Hauptschulen und den Gymnasien wurden die zentralen Tests zunächst nur am Ende der 6. Klasse geschrieben, in der Realschule am Ende der 6. und 8. Klasse. Um schulformübergreifende Auswertungen vornehmen zu können, bot sich deshalb eine Befragung der Lehrkräfte der Klasse 6 an.

3.1 Erhebungsinstrumente

Die Einstellungen gegenüber Vergleichsarbeiten wurden mit den Skalen *allgemeine Akzeptanz zentraler Vergleichsarbeiten* (z.B.: „Zentrale Tests sollten regelmäßig durchgeführt werden."), *Vergleichsarbeiten als Belastung* (z.B.: „Vergleichsarbeiten führen zu Konkurrenz und Missgunst innerhalb der Schulen.") und *Lehrplanvalidität der Vergleichsarbeit* (z.B.: „Die Vergleichsarbeit deckt mit ihren Aufgaben die im Bildungsplan vorgegebenen Lerninhalte und Kompetenzen in diesem Fach gut ab.") gemessen. Die ersten beiden Skalen basieren teilweise auf Items der Skala *Akzeptanz zentraler Testuntersuchungen* von Ditton und Merz (2000). Alle drei Einstellungsskalen bestehen jeweils aus 4-6 Einzelitems mit einem fünfstufigen Likert-Rating. Die internen Konsistenzen (Cronbach's alpha) sind gut und variieren zwischen .83 und .88.

Weitere drei Skalen sind Eigenentwicklungen, die verschiedene Nutzungsoptionen landesweiter Tests abbilden. Für die ersten beiden Skalen wurde auf die Unterscheidung zwischen Selektions- bzw. Auslesediagnostik vs. Modifikations- bzw. Förderdiagnostik bei Horstkemper (2004) zurück gegriffen. Die Skala *förderdiagnostische Nutzung* be-

schreibt, in welchem Maße Vergleichsarbeiten als zusätzliche Unterstützung von Lerndiagnose, Beratung und Förderung gesehen werden können (z.B. „Vergleichsarbeiten sind ein guter Anhaltspunkt, um die Leistung einzelner Schüler einschätzen zu können", „… bieten eine gute Grundlage zur Planung von Fördermaßnahmen für schwächere Schüler."). Inwiefern Vergleichsarbeiten die Benotung und die selektionsdiagnostischen Entscheidungen der Lehrer unterstützen, wird mit der Skala *selektionsdiagnostische Nutzung* erhoben (z.B. „Vergleichsarbeiten tragen auch zur Begründung der Jahresendnote bei.", „… regen zum Nachdenken über den eigenen Bewertungsmaßstab an."). Förder- und selektionsdiagnostische Handlungen lassen sich selbstverständlich in der Praxis nie so idealtypisch trennen, wie dies mit den beiden Skalen versucht wird. Mit den Skalen wird vielmehr der Versuch unternommen, zwei Aspekte diagnostischen Handelns abzubilden, die sich ergänzen aber auch deutlich widersprechen können (Horstkemper 2004).

Mit der Skala *Hinweise für die Unterrichtsgestaltung* wird geprüft, ob Lehrkräfte durch die Leistungsrückmeldungen zu einer Reflexion über ihre inhaltlichen Schwerpunktsetzungen oder die eingesetzten Lernmaterialen bzw. Aufgabenstellungen angeregt werden (z.B. „Die Vergleichsarbeit gibt Hinweise, welche Inhalte in Zukunft verstärkt behandelt werden sollten."). Alle drei Skalen zu den Nutzungsoptionen bestehen aus 5 bis 7 Items und haben gute interne Konsistenzen zwischen .81 und .91.

Als mögliche Prädiktorvariablen auf Ebene der einzelnen Lehrkraft wurde nach Alter, Beschäftigungsumfang und der *Lehrerselbstwirksamkeitserwartung* (Schwarzer/Jerusalem 1999; Schmitz/Schwarzer 2000; alpha =.78) gefragt. Auf Klassenebene wurden folgende Kontextvariablen erfasst: Klassengröße, Anzahl der Schüler mit deutsch als nicht-dominanter Sprache, Notendurchschnitt der Vergleichsarbeit und der Mittelwert der Jahresendnoten in dem Fach, in dem die Vergleichsarbeit geschrieben wurde. Da die Lehrkräfte die Vergleichsarbeiten selbst durchgeführt und ausgewertet haben, sind selbstverständlich gewisse Verzerrungen denkbar. Trotz alledem werden die Vergleichsarbeitsdurchschnitte die Leistung der Klasse objektiver beschreiben als die von den Lehrern angegebenen Notendurchschnitte.

Aus der Differenz der beiden Notendurchschnitte wurde eine weitere Lehrervariable *Tendenz zu strenger Beurteilung* berechnet. Um den institutionellen und regionalen Kontext abzubilden, wurde nach Schulgröße (Anzahl Parallelklassen), Anteil der Schüler mit deutsch als nicht-dominanter Sprache an der Schule, Führungskompetenz der Schulleitung und Schulform gefragt. Die Skala *Führungskompetenz der Schulleitung* stammt aus dem QuaSSU-Projekt (Ditton/Arnoldt/Bornemann 2002) und besteht aus fünf Items mit jeweils vierstufigen Likert-Skalen (z.B. „Die Schulleitung sorgt für einen guten Informationsfluss in der Schule."; alpha = .93)

Um die regionalspezifische Angebotsstruktur im Sekundarbereich abbilden zu können, wurde eine Indexvariable *Attraktivität des regionalen Bildungsangebots* berechnet. Je höher der Wert dieser Variable ist, desto mehr weiterführende Schulformen (Hauptschule, Realschule, Gymnasium und Privatschule) befinden sich am Ort der Schule, an der die befragte Lehrkraft unterrichtet.

3.2 Stichprobe

Um nicht alle baden-württembergischen Schulen anschreiben zu müssen, wurde eine regional stratifizierte Stichprobe von Stadt- und Landkreisen gezogen, in denen sämtliche Hauptschulen, Realschulen und Gymnasien angeschrieben wurden. Die Stichprobe repräsentiert somit Regionen mit sehr hohen, durchschnittlichen und sehr niedrigen gymnasialen Übergangsquoten. Insgesamt wurden 256 öffentliche Schulen mit geschätzten 1050 Klassen in der Jahrgangsstufe 6 angeschrieben (Tabelle 1). Das Schreiben richtete sich an die Schulleitung mit der Bitte, die Fragebögen an die betreffenden Lehrkräfte zu verteilen. Eine genaue Rücklaufstatistik kann nicht angeben werden, liegt aber grob bei ca. einem Drittel, wobei die Beteiligung der Gymnasiallehrer deutlich über der der Haupt- und Realschullehrer liegt.

Tab. 1: **Stichprobe und Rücklauf**

	Angeschriebene Schulen (Anteil an Grundgesamtheit dieser Schulform in BW)	6. Klassen der angeschriebenen Schulen	Rücklauf Fragebögen
HS	161 (13,4 %)	ca. 450	113
RS	53 (12,4 %)	ca. 350	81
GY	42 (11,2 %)	ca. 250	113
	256 (12,7 %)	ca. 1050	307

Anmerkung: Datengrundlage ist die Statistik für das Schuljahr 2004/05 des Statistischen Landesamtes Baden-Württemberg

63,1% der zurückgesandten Fragebögen wurden von Lehrerinnen ausgefüllt, davon 55,6% von Gymnasiallehrerinnen. Die größte Altersgruppe sind die 51- bis 60-Jährigen mit einem Anteil von 41,6% an der Gesamtstichprobe. Vor allem an Realschulen und Gymnasien liegt der Anteil dieser Altersgruppe noch höher. 11,5% der Fragebögen wurden von Lehrkräften unter 30 Jahren ausgefüllt. An Hauptschulen liegt der Anteil dieser jungen Lehrkräfte fast doppelt so hoch bei 20,4%. Die Altersverteilung der Stichprobe entspricht somit in etwa den Altersverteilungen in den einzelnen Schulformen.

38,9% der Befragten sind teilzeitbeschäftigt (Landesdurchschnitt: 41,9%[8]). An Gymnasien ist der Anteil teilzeitbeschäftigter Lehrkräfte in der Stichprobe höher (43,8%), an Hauptschulen niedriger (33,3%). Auch die Verteilung der Schulgröße in der Stichprobe entspricht in etwa den landesweiten Daten. Die befragten Hauptschullehrkräfte unterrichten vor allem an einzügigen (56,6%) oder zweizügigen (35,4%) Schulen, während die Gymnasiallehrkräfte in der Regel an drei- bis fünfzügigen Schulen sind. Die Klassengröße liegt im Durchschnitt bei 24,6 Schülern. An Hauptschulen ist die durchschnittliche Klassengröße dabei wesentlich kleiner (18,3; Landesdurchschnitt:

8 Die angeführten Landesdurchschnittswerte beziehen sich auf das Schuljahr 2005/06. Quelle: Statistisches Landesamt Baden-Württemberg; http://www.statistik.baden-wuerttemberg.de

19,8) als an Realschulen (27,4; Landesdurchschnitt: 26,8) und Gymnasien (28,7; Landesdurchschnitt: 26,2). Dagegen sinkt der Prozentsatz der Schüler mit Deutsch als Zweitsprache von 27,1 an Hauptschulen auf 5,1 an Gymnasien.

In der Hauptschule mussten mindestens zwei Vergleichsarbeiten zu den Kernfächern Deutsch, Mathematik und Englisch von den Schulen ausgewählt werden. In der Realschule und im Gymnasium waren die Tests in den Kernfächern Deutsch und Mathematik verpflichtend. In der Realschule musste zusätzlich eine Vergleichsarbeit in EWG[9] oder Geschichte durchgeführt werden, im Gymnasium in Biologie oder GWG[10]. Die Festlegung der optionalen Vergleichsarbeiten erfolgte auch hier auf Schulebene. Tabelle 2 zeigt, in welchen Fächern an den Schulen der befragten Lehrkräfte Vergleichsarbeiten geschrieben wurden. Nur wenige Gymnasien entschieden sich für eine Vergleichsarbeit im Fächerverbund GWG. An den Realschulen ist das Verhältnis zwischen den Vergleichsarbeiten in den Nebenfächern dagegen ausgewogen. An Hauptschulen wurden vorzugsweise die beiden Hauptfächer Deutsch und Mathematik gewählt. Ein nicht unerheblicher Teil der befragten Hauptschullehrkräfte gab an, dass in allen drei Fächern Vergleichsarbeiten durchgeführt wurden.

Tab. 2: **Vergleichsarbeiten an den Schulen der befragten Lehrkräfte** (in %)

	HS	RS	GY
Deutsch	84,1	97,5	98,2
Mathematik	90,3	96,3	99,1
Englisch	68,1		
EWG		74,1	
Geschichte		72,8	
Bio			79,6
GWG			18,6

Die Vergleichsarbeiten führten generell zu besseren Klassendurchschnittsnoten als die von den Lehrkräften vergebenen durchschnittlichen Zensuren pro Klasse. Diese Differenz wurde auch vom Landesinstitut für Schulentwicklung festgestellt und in einem Bericht[11] an die Schulen zum ersten Pflichteinsatz von Vergleichsarbeiten kritisch thematisiert. Dort findet sich die Erklärung, dass die Pilotierungsschulen weniger Zeit hatten, um die Anforderungen der neuen Bildungspläne umzusetzen, und sich auf den Pilotierungstest auch nicht vorbereiten konnten. Unabhängig von der Plausibilität dieser Erklärung führte das „kritische" Pilotierungsergebnis vermutlich zu einer Reduktion der

9 Fächerverbund Erdkunde – Wirtschaftskunde – Gemeinschaftskunde

10 Fächerverbund Geographie – Wirtschaft – Gemeinschaftskunde

11 Landesinstitut für Schulentwicklung BW (2006): Erfahrungen mit Pilotierung 2005 und Pflichteinsatz 2006. [abgerufen am 21.5.2007 unter http://lbsneu.schule-bw.de/entwicklung/dva/dva_2006/Vergleich06]

Leistungsanforderungen bzw. einer „milderen" Notenumrechungstabelle für den Pflichteinsatz im Jahr 2006.

Um zu prüfen, ob die Stichprobe bezüglich der Lehrerselbstwirksamkeitserwartung als einer zentralen berufsbezogenen Persönlichkeitsvariable verzerrt ist, bot sich ein Vergleich mit den Daten aus dem Modellversuch „Selbstwirksame Schule" an (Schwarzer/Jerusalem 1999). Die Skala wurde dort zu drei Messzeitpunkten im Abstand von einem Jahr eingesetzt und die Mittelwerte der Skalensumme lagen zwischen 28,53 (SD=4,60; n=267) und 29,07 (SD=3,85; n=269). Die hier befragte Lehrergruppe kam auf einen durchschnittlichen Skalensummenwert von 28,43 (SD=4,00; n=298) und entspricht somit hinsichtlich der Lehrerselbstwirksamkeitserwartung den Referenzgruppen bei Schwarzer und Jerusalem.

4. Ergebnisse

Die hier berichteten Mittelwerte beziehen sich nicht auf Einzelitems, sondern jeweils auf den mittleren Skalenwert. Werte über dem semantischen Median von 3 (bei einer 5-stufigen Skala) können als tendenzielle Zustimmung, Werte unter 3 als tendenzielle Ablehnung der Lehrkräfte interpretiert werden.

4.1 Nutzungsmöglichkeiten und Akzeptanz der Vergleichsarbeiten

Wie bewerten Lehrkräfte mögliche Handlungsoptionen im Zusammenhang mit der Ergebnisinterpretation von Vergleichsarbeiten? Die Skala *selektionsdiagnostische Nutzung* wird insgesamt am höchsten eingeschätzt (M=2,82 / SD=0,93) und unterscheidet sich signifikant[12] von den beiden anderen Skalen *förderdiagnostische Nutzung* (M=2,68 / SD=0,91) und *Hinweise für zukünftige Unterrichtsgestaltung* (M=2,71 / SD=0,89). Die Mittelwerte der beiden letztgenannten Skalen unterscheiden sich statistisch nicht voneinander. Wenngleich die Unterschiede zwischen den Skalen substanziell nicht sehr groß sind, zeigt sich im Durchschnitt dennoch eine höhere Affinität zwischen der Interpretation von Vergleichsarbeitsergebnissen und dem unterrichtlichen Handlungsfeld „Notengebung und Selektionsentscheidungen". Alle drei Skalenmittelwerte befinden sich unter dem semantischen Median von 3, d.h. die Nutzungsmöglichkeiten externer Leistungsrückmeldungen werden von Lehrkräften eher verhalten eingeschätzt. Hinter diesen Mittelwerten verbirgt sich ein nicht unerheblicher Teil von Lehrern, die in den bisherigen zentralen Tests keinerlei Nutzen sehen. Da es sich um die erste Runde verpflichtender Vergleichsarbeiten in Baden-Württemberg handelt, darf dieses Ergebnis nicht übermäßig verwundern. Dies gilt ebenso für die drei Skalen zur Einschätzung der Akzeptanz. Auch hier liegen die Mittelwerte im eher negativen Bereich (Allgemeine Ak-

12 T-Tests für gepaarte Stichproben: Selektionsdiagnostische Nutzung – förderdiagnostische Nutzung ($p < 0.01$); selektionsdiagnostische Nutzung – Hinweise für zukünftige Unterrichtsgestaltung ($p < 0.05$).

zeptanz der Vergleichsarbeit: M=2,75; SD=1,04; Vergleichsarbeit als Belastung: M=2,71; SD=0,98; Lehrplanvalidität der Vergleichsarbeit: M=2,78; SD=0,95).

In welchem Maße hängt nun die Einschätzung der Nutzung zentraler Leistungsrückmeldungen von der Akzeptanz und der subjektiv wahrgenommenen Lehrplanvalidität der Tests ab (Forschungsfrage 2)? Hierzu wurden die Akzeptanz- und Nutzungsvariablen miteinander korreliert (Tabelle 4). Die Skala *förderdiagnostische Nutzung* korreliert dabei am höchsten mit der *allgemeinen Akzeptanz zentraler Tests* und der *subjektiv wahrgenommenen Lehrplanvalidität* der Vergleichsarbeit. Die Korrelationen zwischen der *allgemeinen Akzeptanz* und den beiden anderen Nutzungsvariablen ist wesentlich niedriger. Lehrkräfte erwarten von Vergleichsarbeiten somit einen – wie auch immer gearteten – förderdiagnostischen Nutzen. Hinweise auf die zukünftige Unterrichtsgestaltung oder auf die eigene Benotungspraxis hängen nicht in dieser Höhe mit der Akzeptanz zusammen. Dies wird bestätigt, wenn man die negativ formulierte Variable *Vergleichsarbeit als Belastung* betrachtet. Ein zentraler Test wird dann weniger als Belastung wahrgenommen, wenn es den Anschein hat, dass damit Lerndiagnose und Lernförderung unterstützt werden können.

Tab. 3: **Interkorrelationen der Akzeptanz- und Nutzungsvariablen**

	Allgemeine Akzeptanz	VA als Belastung	Lehrplanvalidität der VA	Förderdiagnostische Nutzung	Selektionsdiagnostische Nutzung	Hinweise auf zukünftige Unterrichtsgestaltung
Allgemeine Akzeptanz		-0,51	0,46	0,74	0,58	0,54
VA als Belastung	-0,51		-0,31	-0,48	-0,36	-0,25
Lehrplanvalidität der VA	0,46	-0,31		0,64	0,52	0,45
Förderdiagnostische Nutzung	0,74	-0,48	0,64		0,69	0,59
Selektionsdiagnostische Nutzung	0,58	-0,36	0,52	0,69		0,61
Hinweise auf zukünftige Unterrichtsgestaltung	0,54	-0,25	0,45	0,59	0,61	

Anmerkung: Alle Kriteriumsvariablen korrelieren hoch signifikant miteinander.

Die Einschätzungen zu den drei möglichen Handlungsoptionen korrelieren ebenfalls sehr hoch miteinander. Dies legt die Vermutung nahe, dass Lehrkräfte die verschiedenen Anwendungsbereiche zentraler Vergleichsarbeiten nicht getrennt voneinander sehen.

4.2 Fach- und schulformspezifische Differenzen in der Bewertung von Vergleichsarbeiten

Es ist anzunehmen, dass es schulform- bzw. fachspezifische Differenzen in der Einschätzung der Akzeptanz und der Nutzung zentraler Lernstandserhebungen geben wird

(Forschungsfrage 3). Zur Prüfung dieser Annahme wurden zweifaktorielle Varianzanalysen mit den Faktoren Schulform und Fach berechnet. Um stabile und zwischen den Schulformen besser vergleichbare Ergebnisse zu erhalten, wurden für die nachfolgenden Analysen lediglich die Daten zu den Vergleichsarbeiten in den Hauptfächern Deutsch und Mathematik herangezogen (n=213).

Zunächst wurden die Varianzanalysen der drei Skalen zur Akzeptanzeinschätzung analysiert. Für die Skala *allgemeine Akzeptanz von Vergleichsarbeiten* ergebt sich ein signifikanter Haupteffekt Schulform ($M_{HS} = 3{,}29$; $M_{RS} = 2{,}58$; $M_{GY} = 2{,}47$; $p < 0.001$) und ein Trend für den Haupteffekt Fach ($M_{Deutsch} = 2{,}63$; $M_{Mathe} = 2{,}90$; $p < 0.085$). Die *Skalen Vergleichsarbeiten als Belastung* ($M_{HS} = 2{,}28$; $M_{RS} = 3{,}25$; $M_{GY} = 2{,}86$; $p < 0.001$) und *Lehrplanvalidität der Vergleichsarbeit* ($M_{HS} = 3{,}20$; $M_{RS} = 2{,}52$; $M_{GY} = 2{,}51$; $p < 0.001$) differieren hingegen nur schulformspezifisch. Interaktionseffekte treten nicht auf.

Die schulformspezifischen Ergebnisse lassen sich demnach vor allem auf die Realschul- und Gymnasiallehrer zurückführen. Die Differenzen in der Einschätzung der Vergleichsarbeiten in den einzelnen Fächern werden dabei im Wesentlichen von den schulformspezifischen Differenzen überlagert. Lediglich der tendenziell signifikante Haupteffekt bei der allgemeinen Akzeptanz von Vergleichsarbeiten weist darauf hin, dass der zentrale Test in Deutsch etwas kritischer beurteilt wurde.

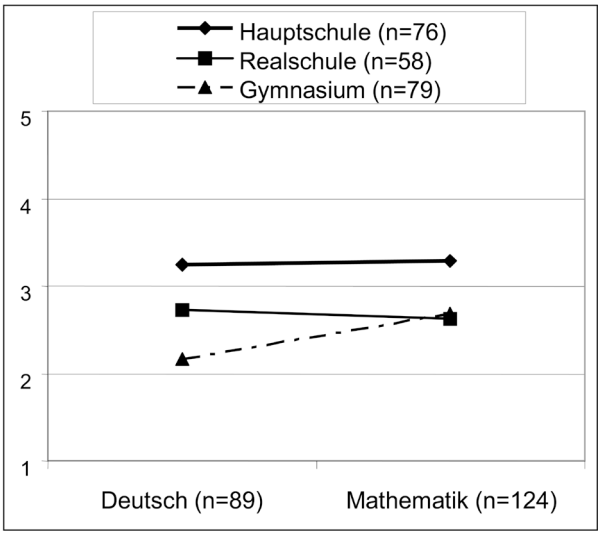

Abb. 1: Selektionsdiagnostische Nutzung von Vergleichsarbeiten nach Schulform und Fach (Anmerkung: Zweifaktorielle Varianzanalyse mit signifikantem Haupteffekt Schulform [p<0.001] und tendenziell signifikantem Interaktionseffekt Schulform x Fach [p=0.06]1)

Bei den Skalen zur Einschätzung der Nutzung von Vergleichsarbeiten gibt es ebenfalls durchgängig einen hoch signifikanten Haupteffekt der Schulform ($p < 0.001$). Dies geht wiederum auf die positive Beurteilung durch Hauptschullehrkräfte zurück. Ein Haupteffekt Schulfach tritt nicht auf. Tendenziell signifikant wird lediglich der Interaktionseffekt „Schulform x Schulfach" bei der selektionsdiagnostischen Nutzung (Abbildung 1).

4.3 Einflüsse von Kontextvariablen auf Akzeptanz und subjektiv eingeschätzte Nutzung

Mithilfe von Regressionsanalysen soll abschließend der Frage nachgegangen werden, welche individuellen, klassenspezifischen und schulspezifischen Kontextvariablen die Sichtweise der Lehrkräfte auf die Nutzung zentraler Lernstandserhebungen moderieren (Forschungsfrage 4). Hierzu wurden die möglichen Prädiktorvariablen in drei Ebenen eingeteilt. Auf Ebene 1 befinden sich die individuellen Lehrermerkmale Alter und Lehrerselbstwirksamkeit. Aus der Differenz zwischen Vergleichsarbeitsdurchschnitt und Notendurchschnitt der Klasse in dem jeweiligen Fach wurde eine weitere Lehrervariable *Tendenz zur strengen Benotung* berechnet (Notendurchschnitt – Vergleichsarbeitsdurchschnitt).

Die Angaben zur Klasse, in der die Vergleichsarbeit durchgeführt wurde, bilden Ebene 2: Klassengröße, Anteil der Schüler mit deutsch als nicht dominanter Sprache und der Notendurchschnitt in der Vergleichsarbeit. Der von den Lehrkräften angegebene Notendurchschnitt in der Vergleichsarbeit wurde als objektiv gemessenes und vergleichbares Leistungskriterium interpretiert.

Um den institutionellen und regionalen Kontext der Schule beschreiben zu können, werden folgende Merkmale auf Ebene 3 in die Analysen mit einbezogen: Schulform, Schulgröße, geschätzter Schüleranteil mit deutsch als nicht dominanter Sprache, Führungskompetenz der Schulleitung, Attraktivität des regionalen Bildungsangebots und Region (nach gymnasialen Übergangsquoten).

Um die Bedeutung der einzelnen Prädiktorvariablen auf den unterschiedlichen Ebenen abschätzen zu können, wurden diese zuerst mit den 6 Kriteriumsvariablen korreliert (Tabelle 4, S. 110). Die Korrelationsmatrix zeigt signifikante und substanziell bedeutsame Korrelationen mit den Kriteriumsvariablen in allen drei Bereichen. Lediglich das Alter der Lehrkräfte und die regionale Zuordnung spielen keine Rolle. Die wichtigsten Kontextvariablen zur Erklärung der Beurteilung zentraler Vergleichsarbeiten befinden sich allerdings auf der Ebene der Schulklasse und der Einzelschule.

Da Wechselwirkungen und Substitutionseffekte zwischen den einzelnen Prädiktorvariablen sehr wahrscheinlich sind (z.B. Migrantenanteil und Schulform), wurden Regressionsanalysen durchgeführt, um die Interpretation der Zusammenhänge auf eine stabilere Grundlage zu stellen. Bis auf den *Notendurchschnitt im Zeugnis* gehen alle in der Korrelationstabelle aufgelisteten Prädiktorvariablen in die Regressionsrechnung ein. Die Varianz des Notendurchschnitts im Zeugnis ist bereits in der Variable *Tendenz zur strengen Benotung* enthalten und würde zu Multikolinearität führen.

Die Regressionsanalysen führen zu einer Konzentration der bisher ermittelten Zusammenhänge auf wesentliche Kontextvariablen (Tabelle 5). Dabei zeigt sich, dass die in den Korrelationstabellen noch signifikanten Zusammenhänge mit Schulgröße, Migrantenanteil und Führungskompetenz der Schulleitung in der Regressionsrechnung ihre Signifikanz verlieren. Die bivariaten Korrelationen lassen sich vermutlich komplett auf die Schulform als zentrale Prädiktorvariable auf Ebene 3 zurückführen. Die Attraktivität des regionalen Bildungsangebots entpuppt sich vor allem als institutionelle Prädiktorvariable für die selektionsdiagnostische Nutzung der Vergleichsarbeiten. Dies würde

zu der spekulativen These führen, dass sich regionale Bildungsmärkte unter Umständen auf die Selektionspraxis und die Notengebung an Schulen auswirken können. Einen gewissen regionalen Einfluss auf die Wahrnehmung von zentralen Tests scheint es ebenfalls zu geben. Vor allem in städtischen Regionen werden Vergleichsarbeiten eher als Belastung erlebt als in überwiegend ländlichen Gebieten.

Tab. 4: **Korrelationen zwischen Kriteriumsvariablen und Prädiktorvariablen**

	Allgemeine Akzeptanz von VA	VA als Belastung	Lehrplanvalidität der VA	VA unterstützt Lerndiagnose, Beratung und Förderung	VA unterstützt Benotung und Selektionsentscheidungen	VA gibt Hinweise auf Unterrichtsgestaltung
Lehrkraft						
Alter						
Beschäftigungsumfang						
Lehrerselbstwirksamkeitserwartung			.15 *	.13 *		
Tendenz zur Strenge			-.20 **	-.14 *		
Schulklasse						
Klassengröße	-.31 ***	.29 ***	-.35 ***	-.40 ***	-.36 ***	-.35 ***
Anteil Schüler mit deutsch als nicht dominanter Sprache	.29 ***	-.27 ***	.29 ***	.31 ***	.20 **	.16 *
Notendurchschnitt in der Vergleichsarbeit	.16 *			.26 ***	.18 **	.20 **
Notendurchschnitt im Zeugnis	.16 *	-.18 **		0,17 **		.18 **
Schule						
Schulgröße (Anzahl Parallelklassen in Jahrgangsstufe 6)	-.30 ***	.26 ***	-.29 ***	-.36 ***	-.32 ***	-.26 ***
Geschätzter Anteil Schüler mit deutsch als nicht dominanter Sprache	.22 ***	-.19 **	.19 **	.26 ***	.17 **	.14 *
Führungskompetenz Schulleitung			.16 *	.13 *	.16 **	
Region (gymnasiale Übergangsquote)						
Attraktivität des regionalen Bildungsangebots	-.14 *		-.13 *	-.13 *		

Anmerkungen zum Signifikanzniveau: * = $p < 0{,}05$; ** = $p < 0{,}01$; *** = $p < 0{,}001$. Nicht signifikante Korrelationen wurden ausgeblendet.

Tab. 5: **Regression der Vergleichsarbeitsbewertungen durch Lehrkräfte auf individuelle, klassenbezogene und institutionelle Kontextmerkmale (beta-Gewichte)**

		Allgemeine Akzeptanz	VA als Belastung	Lehrplanvalidität	förderdiagnostische Nutzung	selektionsdiagnostische Nutzung	Hinweise auf Unterricht
	Alter			0,17 *		0,14 *	
	Beschäftigungsumfang						
	Lehrerselbstwirksamkeitserwartung						
	Tendenz zu strenger Beurteilung			-0,38 ***			
	Klassengröße		0,22 *	-0,25 **	-0,20 *	-0,26 **	-0,25 **
	Anteil Schüler mit deutsch als nicht dominanter Sprache	.24 *		0,28 **	0,20 *		
	Notendurchschnitt in der Vergleichsarbeit			-0,34 **			
	Schulgröße						
	Geschätzter Anteil Schüler mit deutsch als nicht dominanter Sprache						
	Führungskompetenz Schulleitung						
	Schulform	-.32 **			-0,41 ***	-0,23 *	-0,30 *
	Attraktivität des regionalen Bildungsangebots					0,19 **	
	Region (gymnasiale Übergangsquote)		0,17 *				
R^2		**.265**	**.243**	**.280**	**.322**	**.226**	**.188**
n		188	188	187	188	188	188

Anmerkungen zum Signifikanzniveau: * = p < 0,05; ** = p < 0,01; *** = p < 0,001. Nicht signifikante beta-Werte wurden ausgeblendet

Alle drei klassenbezogenen Prädiktorvariablen spielen bei der Vorhersage der Kriteriumsvariablen eine gewisse Rolle. Lehrkräfte mit größeren Klassen sehen Akzeptanz und Nutzung zentraler Tests generell kritischer. Ob dies auf den höheren Korrekturaufwand bei der Auswertung der Vergleichsarbeit zurückgeführt werden kann oder ob Lehrkräfte mit großen Klassen durch eine insgesamt höhere Arbeitsbelastung Innovationen eher skeptisch beurteilen, lässt sich mit den hier vorliegenden Daten nicht entscheiden. Dagegen bestätigt sich der positive Einfluss eines hohen Migrantenanteils in der Klasse auf die Einschätzung der allgemeinen Akzeptanz, der wahrgenommenen Lehrplanvalidität und der förderdiagnostischen Nutzung von Vergleichsarbeiten. Es scheint, dass Lehrkräfte in Klassen mit heterogener Lernausgangslage an objektiven Diagnoseinstrumen-

ten interessiert sind. Erwartungsgemäß ist, dass der Notendurchschnitt in der Vergleichsarbeit negativ mit der Einschätzung der Lehrplanvalidität zusammenhängt. Je besser die Klasse abschneidet, desto eher wird von den Lehrkräften die Passung zwischen Test und Lehrplan bestätigt.

Auf individueller Ebene finden sich nur wenige signifikante Prädiktoren. Ältere Lehrkräfte schätzen die Lehrplanvalidität und die selektionsdiagnostische Nutzung der Vergleichsarbeiten höher ein als ihre jüngeren Kollegen. Die Lehrerselbstwirksamkeit hat keinen Effekt auf die Einschätzung zentraler Tests. Dafür ist die Tendenz zur strengen Beurteilung ein starker negativer Prädiktor für die von den Lehrkräften wahrgenommene Lehrplanvalidität der Vergleichsarbeit. Es ist anzunehmen, dass sich dahinter vor allem eine Kritik an den relativ leichten Aufgabenstellungen verbirgt. Lehrkräfte mit strengen Benotungsmaßstäben hätten gerne schwierige Tests, die ihre eigenen Maßstäbe und Ansprüche dann auch bestätigen könnten.

5. Zusammenfassung und Diskussion

Die Implementation einer datenbasierten Schul- und Unterrichtsentwicklung ist eine langwierige Angelegenheit und es darf nicht verwundern, wenn die Mehrheit der hier befragten Lehrkräfte von einer pädagogischen Nutzung zentraler Leistungsrückmeldungen nicht überzeugt ist. Umso wichtiger ist es, die schulischen Prozesse und „Sensibilitäten" herauszuarbeiten, wenn Praktiker vor Ort mit Vergleichsarbeiten konfrontiert werden. Beispielsweise können Rückmeldedaten eher Hinweise für die eigene Notengebung und Selektionsdiagnostik geben, während die Nutzung im Rahmen des förderdiagnostischen Handelns oder der Reflexion über Unterricht noch ausbaufähig ist. Dieser Befund korrespondiert im weitesten Sinne mit den Evaluationsergebnissen des Deutschfreiburger Übergangsmodells (Baeriswyl/Wandeler/Trautwein/Oswald 2006) und der Befragung zur Nutzung von Klassencockpit im Kanton Zürich von Moser (2003). Im ersten Fall wurde zum Beispiel ein zentrales Testverfahren ganz gezielt zur Optimierung der selektionsdiagnostischen Entscheidung am Ende der Primarstufe eingesetzt.

Gleichzeitig muss beachtet werden, dass Akzeptanz und Einschätzung der Nutzung zentraler Tests eng miteinander zusammenhängen. Die Rückmeldungen von Vergleichsarbeiten werden nur dann von Lehrkräften dauerhaft als wichtige, handlungsleitende Informationen akzeptiert, wenn auch eine substanzielle, förderdiagnostische Nutzung dieser Daten erkennbar ist. Ein Großteil der Lehrkräfte scheint davon nicht überzeugt zu sein, zumal durch die bisher stark summative Ausrichtung des Tests (am Ende eines zweijährigen Bildungsabschnitts) eine Ableitung differenzierter Lernstandsbeschreibungen nicht möglich erscheint.

Besonders auffallend sind die schulformspezifischen Differenzen in der Wahrnehmung der Vergleichsarbeiten. Bei einer Weiterentwicklung der Lernstandserhebungen wird eine verstärkte Orientierung an den Bedarfen der einzelnen Schulformen unabdingbar sein. Ebenso müssen die Stützsysteme zur Implementation einer ergebnisorien-

tierten Unterrichtsentwicklung (z.B. Lehrerfortbildung) wesentlich stärker auf kritische Vorbehalte und Wissensdefizite der Lehrkräfte eingehen (vgl. O'Day 2004).

Dass man an den Schulen noch weit von einer systematischen Strategie zur Nutzung externer Leistungsrückmeldungen entfernt ist, legen die Befunde zum Einfluss der Kontextvariablen nahe. Die Nutzung der Daten ist bisher eher die „Privatsache" einer einzelnen Lehrkraft. Deswegen spielen vermutlich institutionelle Faktoren, wie z.B. die Führungskompetenz der Schulleitung keine Rolle bei der Einschätzung verschiedener Nutzungsoptionen. Schwer zu erklären ist dagegen der Einfluss des regionalen Bildungsangebots auf die Wahrnehmung der selektionsdiagnostischen Nutzung zentraler Tests. Es wäre denkbar, dass sich die Attraktivität des Bildungsangebots über die Bildungsaspiration der Eltern auf die Wahrnehmung und Bedeutung von selektionsdiagnostischen Entscheidungen innerhalb einer Schule auswirken könnte.

Auf individueller Ebene kann mit den erhobenen Kontextvariablen ebenfalls nur ein geringer Varianzanteil geklärt werden. Erwartungswidrig ist vor allem, dass die Lehrerselbstwirksamkeitserwartung keinerlei Einfluss auf die Akzeptanz oder Nutzung der Rückmeldedaten hat. Im entsprechenden Zyklenmodell von Helmke und Hosenfeld (2005) sowie in der *feedback intervention theory* bei Kluger und DeNisi (1996) wird diese Variable als wichtige individuelle Kontextbedingung für die Nutzung von Rückmeldungen aufgeführt. Dies könnte damit zusammenhängen, dass in der hier vorliegenden Studie lediglich nach Nutzungsmöglichkeiten gefragt wurde. In den tatsächlichen Handlungssituationen werden berufsbezogene Persönlichkeitsmerkmale vermutliche eine größere Rolle spielen.

Die hier berichtete Studie gibt erste Einblicke in die Art und Weise, wie Lehrkräfte Vergleichsarbeitsdaten rezipieren und zu nutzen gedenken. Wie in einem relativ neuen Forschungsfeld nicht anders zu erwarten, werfen die Ergebnisse mehr neue Fragen auf als beantwortet werden können. Es ist vor allem unklar, ob die mit dem Fragebogen erfassten Nutzungsoptionen von Vergleichsarbeiten auch mit dem tatsächlichen evaluativen Handeln der Lehrkräfte übereinstimmen. Mit rein quantitativen Befragungsdaten wird man dabei zu keiner befriedigenden Antwort kommen können. Für die Beschreibung schulinterner Evaluations- und Entwicklungsprozesse werden auch qualitative Zugänge nötig sein.

Die zentrale Fragestellung wird weiterhin sein, ob es an den Schulen gelingt, die Interpretation externer Leistungsrückmeldungen sinnvoll in innerschulische Evaluationsprozesse einbinden zu können. Dies wird nur gelingen können, wenn sich die in den offiziellen Gremien praktizierte Dateninterpretation an den realen Erfordernissen der Lehrkräfte orientieren wird. Wenn Vergleichsarbeitsdaten auf Schulebene allerdings nur symbolisch genutzt werden (Rossi/Freeman 1993), um bei einer externen Evaluation die Außendarstellung zu optimieren, wird sich die von den Lehrkräften praktizierte Dateninterpretation und Datennutzung entweder von den institutionellen Prozessen abkoppeln oder es findet auf Lehrerebene ebenfalls eine nur symbolische Nutzung von Testergebnissen statt. In weiteren Studien wird es vor allem darauf ankommen, weitere Gelingensbedingungen für eine direkte, instrumentelle Nutzung von Rückmeldeinformationen auf allen Ebenen der Einzelschule zu beschreiben.

Literatur

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

Baumert, J. (2001): Vergleichende Leistungsmessung im Bildungsbereich. In: Zeitschrift für Pädagogik, 43. Beiheft. Zukunftsfragen der Bildung, S. 13–36.

Böttcher, W./ Holtappels, H.-G./ Brohm, M. (Hrsg.) (2006): Evaluation im Bildungswesen – Eine Einführung in Grundlagen und Praxisbeispiele. Weinheim und München: Juventa.

Ditton, H./Arnoldt, B./Bornemann, E. (2002): Entwicklung und Implementation eines extern unterstützten Systems der Qualitätssicherung an Schulen – QuaSS U. In: Zeitschrift für Pädagogik, 45. Beiheft, S. 374–389.

Ditton, H./Merz, D. (2000): Qualität von Schule und Unterricht – Kurzbericht über erste Ergebnisse einer Untersuchung an bayerischen Schulen. Katholische Universität Eichstätt / Universität Osnabrück (3.1.2003 unter http://www.quassu.net/index.htm).

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. S. 19–40.

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

Helmke, A./Hosenfeld, I./Schrader, F.-W. (2004): Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In: Arnold, R./Griese, C. (Hrsg.): Schulleitung und Schulentwicklung. Hohengehren: Schneider-Verlag. S. 119–144.

Horstkemper M. (2004): Diagnosekompetenz als Teil pädagogischer Professionalität. In: Neue Sammlung, 44, H. 2, S. 201–214.

Imhof, M. (2005): Zur Rezeption der Ergebnisse der PISA-Studie durch Lehrer und Lehrerinnen. In: Unterrichtswissenschaft 33, H. 3, S. 255–271.

Klieme, E (2004): Begründung, Implementation und Wirkung von Bildungsstandards: Aktuelle Diskussionslinien und empirische Befunde. In: Zeitschrift für Pädagogik 50, H. 5, S. 625–634.

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.

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

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

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

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 http://www.lehrmittelverlag.ch/downloads/dateien/ Evaluation%20Klassencockpit.pdf]

O'Day, J. A. (2002): Complexity, accountability, and school improvement. Harvard Educational Review 72(3), p 293–329.

O'Day, J. A. (2004): Complexity, Accountability, and School Improvement. In: Fuhrman, S.H./ Elmore, R.F. (Eds.): Redesigning Accountability Systems for Education. New York/London: Teachers College Press. p. 15–43.

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

Peek, R./Dobbelstein, P. (2006a): Benchmarks als Input für die Schulentwicklung – das Beispiel der Lernstandserhebungen in Nordrhein-Westfalen. In Kuper H./Schneewind, J. (Hrsg.): Rückmeldung und Rezeption von Forschungsergebnissen. New York, München, Berlin: Waxmann. S. 41–58.

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

Rolff, H.-G. (2001): Was bringt die vergleichende Leistungsmessung für die pädagogische Arbeit an Schulen? In: Weinert, F.E. (Hrsg.): Leistungsmessungen in Schulen. Weinheim, Basel: Beltz. S. 337–365.

Rossi, P.H./Freeman, H.E. (1993): Evaluation: A systematic approach. London: Sage.

Scheerens, J./Glas, C./Thomas, S.M. (2003): Educational evaluation, assessment, and monitoring – a systemic approach. Lisse: Swets & Zeitlinger.

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

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

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

Sedlmeier, P./ Böhm, M./ Lindner, S./Schmidt, M. (2006): PISA aus Lehrersicht. Mögliche Ursachen und Verbesserungsvorschläge. In: Unterrichtswissenschaft 34, H. 1, S. 46–69.

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

Weiss, C. H. (1998): Improving the use of evaluations: Whose job is it anyway? In: Reynolds, A.J./Walberg, H.J. (Eds.): Advances in educational productivity, Vol. 7, Greenwich/London: JAI Press. pp. 263–276.

Abstract: *Standardized tests are meant to initiate and support data-based development of schooling and instruction at individual schools. This, however, presupposes the acceptance of central tests by the teaching staff and the targeted use of achievement feedback for specific areas of instructional acting. In a study carried out right after the first round of obligatory standardized tests in Baden-Württemberg, teachers (n=307) were asked to respond to a questionnaire concerning the pedagogical value of feedback on achievement data and the possible effect of context factors. The results mainly point to school-type-specific differences in the perception of standardized tests. Furthermore, teachers see the value of achievement data rather in their use for Professional judgement and selection, although the acceptance of standardized tests is closely related to their use for the diagnosis of special needs.*

Anschrift des Autors:
Dr. Uwe Maier, Institut für Erziehungswissenschaft, Päd. Hochschule Schwäbisch Gmünd, Oberbettringerstraße 200, 73525 Schwäbisch Gmünd.

Anhang

Items und Skalen zur Erfassung der Rezeption und Nutzung von Vergleichsarbeiten

Allgemeine Akzeptanz von Vergleichsarbeiten (alpha = .88; n=284)

Die Vergleichsarbeit:
- sollte regelmäßig durchgeführt werden.
- ist für die Arbeit der Schulen sehr wichtig.
- trägt dazu bei, dass man sich in den Schulen mehr bemüht.
- gibt eine objektive Basis ab, um zu sehen, wo eine Schule steht.
- ist eine wichtige Grundlage, um Unterricht weiterentwickeln zu können.
- nützt für meine eigentliche Arbeit als Lehrer wenig (-).

Vergleichsarbeiten als Belastung (alpha = .83; n=283)

Die Vergleichsarbeit:
- führt zu Konkurrenz und Missgunst innerhalb der Schulen.
- übt zusätzlichen Druck auf Schulen und Lehrer aus.
- führt dazu, dass nur noch für den Test geübt wird.
- bringt Unruhe in die Schulen.
- schafft mehr Probleme als sie nützt.

Lehrplanvalidität der Vergleichsarbeiten (alpha = .85; n=294)

Die Vergleichsarbeit:
- deckt mit ihren Aufgaben die im Bildungsplan vorgegebenen Lerninhalte und Kompetenzen in diesem Fach gut ab.
- stimmt in ihren Teilbereichen mit der Gewichtung der Lerninhalte und Kompetenzen im Bildungsplan überein.
- hat eine Punkteverteilung, die der Gewichtung der Lerninhalte und Kompetenzen im Bildungsplan entspricht.
- entspricht dem im Bildungsplan geforderten Leistungsniveau.

VA unterstützt Lerndiagnose, Beratung und Förderung (alpha = .91; n=294)

Die Ergebnisse der Vergleichsarbeit:
- sind ein guter Anhaltspunkt, um die Leistung einzelner Schüler einschätzen zu können.
- bieten eine gute Grundlage zur Planung von Fördermaßnahmen für schwächere Schüler.
- sind eine tragfähige Argumentationsbasis für Beratungsgespräche mit Eltern.
- helfen mir, die Stärken und Schwächen der Schüler deutlicher benennen zu können.

- geben mir zusätzliche diagnostische Hinweise.
- sind eine gute Gesamtbeurteilung der Schülerleistung.
- spiegeln die Leistungsfähigkeit meiner Schüler nicht so gut wider, wie meine eigenen Klassenarbeiten. (-)

VA unterstützt Benotung und Selektionsentscheidungen (alpha = .81; n=297)

Die Ergebnisse der Vergleichsarbeit:
- tragen auch zur Begründung der Jahresendnote bei.
- in die Jahresendnote mit einfließen zu lassen ist sinnvoll.
- regen zum Nachdenken über den eigenen Bewertungsmaßstab an.
- geben Hinweise darauf, ob die eigenen Klassenarbeiten zu schwer oder zu leicht sind.
- sind bei unsicheren Versetzungsentscheidungen ein zusätzliches Entscheidungskriterium.

VA gibt Hinweise für zukünftige Unterrichtsgestaltung (alpha = .88; n=299)

Die Vergleichsarbeit gibt mir für dieses Fach zusätzliche Hinweise:
- welche Inhalte in Zukunft verstärkt behandelt werden sollten.
- welche Aufgabenstellungen in Zukunft besser geübt werden müssen.
- welche Lerninhalte häufiger wiederholt werden müssen.
- ob die Reihenfolge der behandelten Stoffgebiete geändert werden sollte.
- welche neuen Aufgabenstellungen ich in meinen Unterricht einbinden muss.
- welche Stoffgebiete neu aufgenommen werden sollten.
- welche Lernmaterialien (z.B. Bücher, Arbeitshefte,...) für meinen zukünftigen Unterricht gut geeignet sind.