Evaluation – und was danach? Ergebnisse der Schulleiterbefragung im Rahmen der Rezeptionsstudie WALZER[1]

Friedrich-Wilhelm Schrader und Andreas Helmke

In einer Schulleiterbefragung wurde untersucht, wie die den Schulen zurückgemeldeten Leistungsergebnisse einer landesweiten Evaluationsstudie (MARKUS-Studie in Rheinland-Pfalz) in den Schulen rezipiert werden und zu welchen Veränderungen sie führen. Dazu wurden Schulleiter (N = 52) zu verschiedenen Aspekten dieser Rückmeldungen, zu allgemeinen Einschätzungen und Bewertungen und verschiedenen Aktivitäten der Qualitätssicherung befragt. Dabei zeigt sich, dass die zurückgemeldeten Ergebnisse von der Mehrzahl der Schulen als Hinweise auf notwendige Änderungen gesehen werden und an den Schulen verschiedene Aktivitäten des Qualitätsmanagements zur Folge haben. Weitere Analysen zeigen, dass sich der Einsatz von qualitätssichernden Massnahmen durch verschiedene Einschätzungen und Einstellungen der Schulleiter/innen vorhersagen lässt.

Einleitung

Spätestens seit der Publikation der Ergebnisse von PISA 2000 ist klar, dass die regelmässige Überwachung («monitoring») der Effizienz des Bildungssystems unabdingbar ist, weil ohne eine fundierte empirische Datenbasis keine begründeten Konsequenzen für die Beseitigung von Schwachstellen – oder auch den Ausbau von Stärken – und für die Verbesserung von Schule und Unterricht gezogen werden können. Dies wird keineswegs nur in Deutschland so gesehen, sondern beispielsweise auch in der Schweiz, die bei PISA 2000 lediglich mit den Schülerleistungen im Fach Mathematik zufrieden sein konnte und weniger mit den Leistungen in den Bereichen Naturwissenschaften und Leseverständnis. Auch in der öffentlichen Diskussion, wie sie in den Schweizer Medien stattfand, wird dies deutlich – siehe die Dokumentation der OECD «PISA in the news in Switzerland» (2001-2002).

Angesichts der überragenden Bedeutung und der erheblichen Kosten umfassender Evaluationsstudien auf internationaler Ebene (inbesondere die TIMSS- und PISA-Zyklen), in nationalem Rahmen (wie das PISA-Zusatzprojekt DESI

der KMK in Deutschland; Beck & Klieme, in Druck; Helmke, Goebel, Hosenfeld, Schrader, Vo & Wagner, in Druck) sowie auf regionaler (Bundesländer, Kantone) Ebene, wie die Primarstufenuntersuchungen von Moser (Moser & Rhyn, 2000) im Kanton Zürich sowie die grossen Länderstudien QuaSUM in Brandenburg (Lehmann, Peek, Gänsfuss, Lutkat, Mücke & Barth, 2000b), LAU in Hamburg (Lehmann, Husfeldt & Peek, 2001) und MARKUS (Helmke & Jäger, 2002) in Rheinland-Pfalz, ist es dringend nötig, die Frage nach den Zielen solcher Evaluationsstudien nicht ganz aus den Augen zu verlieren.

Die unbestritten primäre Funktion aller dieser Evaluationsstudien ist die *Standortbestimmung*, d.h. die vergleichende Orientierung an durchschnittlichen Werten, an den Ergebnissen der Besseren und Besten («benchmarking») und an sachlichen Kriterien, z.B. dem prozentualen Anteil der Schüler, die bestimmte wohldefinierte Kompetenzstufen erreicht oder verfehlt haben. Neben dieses «system monitoring» tritt jedoch eine zweite Komponente der Evaluation, die «improvement»-Komponente: Evaluation durch vergleichende Leistungsstudien ist kein Selbstzweck, sondern soll ja letztendlich der *Verbesserung* von Schule und Unterricht, Lehren und Lernen, fachlichen und überfachlichen Kompetenzen der Schülerinnen und Schüler dienen (Helmke, 2000).

Floskeln vom Typ «Nicht vermessen, sondern entwickeln» oder «Die Sau wird vom Wiegen nicht fetter» betonen allerdings einseitig die zweite Komponente – die Verbesserung – bei gleichzeitiger Geringschätzung oder Ausserachtlassung der monitoring-Komponente. Die damit verbundene, in der Vergangenheit insbesondere in Deutschland stark vertretene empirie- und evaluationsfeindliche Orientierung in der Pädagogik und in der Schulpraxis hat mit dazu beigetragen, dass sich Deutschland aus den internationalen Vergleichsstudien lange Zeit so gut wie vollständig herausgehalten hat und dass bundesweite Evaluationsstudien (so geschehen in Hamburg und Berlin bei PISA-E 2000) behindert wurden – mit dem bekannten Ergebnis, dass diese beiden Stadtstaaten infolge zu geringer Stichprobenquoten aus den Analysen des Bundesländervergleichs (Studie PISA-E) ausgeschlossen werden mussten (Baumert, Artelt, Carstensen, Sibberns & Stanat, 2002).

Wir wollen nicht verhehlen, dass in der Schweiz diesbezüglich differenzierter diskutiert und publiziert wurde. Um ein Beispiel zu nennen: «Natürlich ist die Bauernregel, dass eine Sau vom Wiegen nicht fetter werde, nur der eine Teil der Wahrheit. Denn ganz ohne Wiegen gehts auch nicht. Was der Spruch im Kern meint: Wer sich zu fest aufs Wiegen verlegt, versäumt möglicherweise das, worauf es ankommt, nämlich die gute Aufzucht selbst» (Anton Strittmatter in der «Weltwoche», 27/2002).

Es wäre jedoch naiv zu glauben, dass Monitoring und Evaluation gleichsam automatisch zu Verbesserungen der Lehr-Lern-Situation führen würden. Im Zusammenhang mit den zahlreichen aktuellen und künftigen Evaluationsstudien wird daher ein Forschungstyp grosses Gewicht erhalten, der bisher nur in Umrissen erkennbar ist: *Rezeptionsstudien*, d.h. Studien zu den Bedingungen und

zum Verlauf der Rezeption von Evaluationsergebnissen und ihrer Transformation in Massnahmen zur Verbesserung des Lehrens und Lernens. Der relative Stellenwert dieses Forschungsprogramms lässt sich am besten in Form einer Abbildung veranschaulichen (vgl. Abbildung 1):

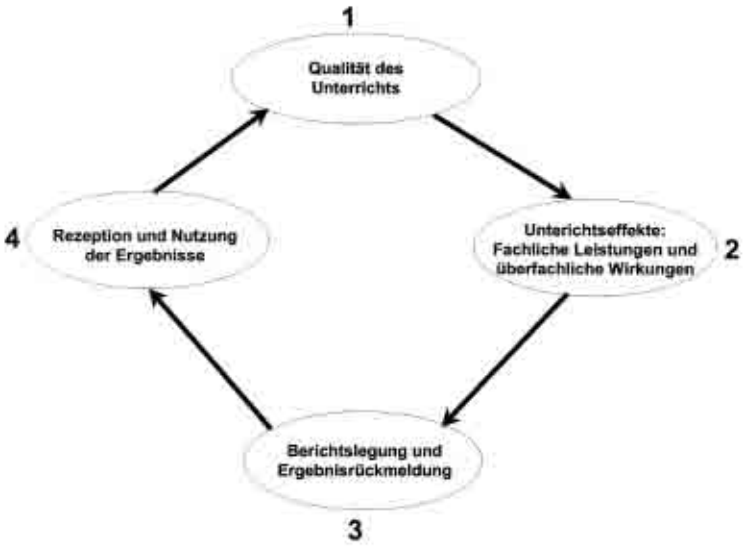

Abbildung 1: Die Verortung des Forschungsprogramms zur Rezeption von Evaluation

Unterrichtsforschung und Lehr-Lern-Forschung sind an den Stationen 1 und 2 des obigen Zyklus angesiedelt: Dazu gehören Untersuchungen in der Tradition des Prozess-Produkt-Paradigmas zur Unterrichtsqualität und -wirksamkeit wie auch die Large Scale Survey- und Längsschnittstudien, sofern sie Bedingungen und Folgen des Unterrichts thematisieren. Die grossen Evaluationsstudien wie PISA oder TIMSS dagegen sind an Station 2 anzusiedeln, weil es bei ihnen schwerpunktmässig um den *Ertrag* von Schule und Unterricht und weniger um die systematische Analyse der Bedingungen geht. Wie man – unter Nutzung kognitions-, motivations- und sozialpsychologischer Erkenntnisse – Berichte und Ergebnisrückmeldungen gestaltet, betrifft den Schritt von 2 zu 3. Hier gibt es nahezu keine wissenschaftliche Basis, keine methodischen Standards; hier wird probiert (und kopiert). Das interessantere Defizit betrifft dagegen die Strecke 3-4: Wie werden Studien vom Typ PISA, TIMSS, MARKUS rezipiert – von den direkt Betroffenen, deren Klassen/Schulen untersucht wurden und die mit Schul- oder Klassenrückmeldungen konfrontiert werden, und darüber hinaus von Lehrkräften, Schulleitungen und Schulpolitikern insgesamt? Noch bri-

santer die Folgefrage: Was folgt aus der wie auch immer gearteten Rezeption für die Unterrichtspraxis: Haben die rückgemeldeten Ergebnisse einen nachweislichen Effekt auf Schule und Unterricht (4-1), und bewirken sie letztendlich verbessertes Lernen, gesteigerte Kompetenzen und solideres Fachwissen auf Seiten der Zielgruppe, der Schülerinnen und Schüler?

Zum Stand der Rezeptionsforschung

Wie Stamm (2002) in einem Übersichtsartikel mit dem vielsagenden Titel «Evaluation und ihre Folgen: Eine unterschätzte pädagogische Herausforderung» schreibt, hat die erziehungswissenschaftliche Forschung zur Frage der Rezeption bisher keinen nennenswerten Beitrag geleistet, «so dass Fragen, beispielsweise, was mit den zur Verfügung gestellten Evaluationsergebnissen geschieht, [...] ob dieses Wissen richtig oder falsch oder überhaupt rezipiert und adäquat weiterverwendet wird, aus pädagogischen Gründen nahezu unbeantwortet geblieben sind» (S. 2). Hier sind uns neben dem BMBF-Projekt von Klemm («Zur Nutzung grossflächiger Tests für die Schulentwicklung. Exemplarische Analyse der Erfahrungen aus England, Frankreich und den Niederlanden»; vgl. auch v. Ackeren, 2002; v. Ackeren & Klemm, 2000), den Arbeiten von Klemm & Schratz (2002) und Rolff (2001) an empirischen Arbeiten nur diejenige von Kohler (2002) zur Rezeption von **TIMSS** bekannt. Dort wurden Lehrkräfte unter anderem auch nach der Bereitschaft gefragt, den eigenen Unterricht infolge der TIMSS-Ergebnisse zu ändern und an weiteren Evaluationsstudien teilzunehmen. Allerdings ging es hier nicht darum, Wirkungen von Ergebnisrückmeldungen an Schulen und Klassen zu verfolgen, sondern um ganz generelle Einschätzungen, die unabhängig von der eigenen Teilnahme an der Studie sind.

Spezifische Ergebnisrückmeldungen an die Schulen erfolgten im Rahmen einer sich an TIMSS anschliessenden Evaluationsstudie von Klieme, Baumert und Schwippert (2000). Da die Initiative zur Evaluation hier von den teilnehmenden Schulen selbst ausging, ist diese Untersuchung für das Verständnis der Rezeption von Ergebnissen aus gross angelegten Schulleistungsstudien allerdings nur bedingt aussagekräftig. Systematische klassen- bzw. kursbezogene Rückmeldungen an die beteiligten Schulen mit dem Ziel, Prozesse der Schul- und Unterrichtsentwicklung in Gang zu setzen, waren Bestandteil der Hamburger Schuluntersuchung **LAU** (**L**ern-**A**usgangslagen-**U**ntersuchung), an der alle Fünftklässler eines Schuljahres teilnahmen und bei der Lernleistungen und schulbezogene Einstellungen vom Ende der Grundschulzeit bis zur 11. Klassenstufe untersucht wurden (Lehmann et al., 2001; Lehmann, Peek & Gänsfuss, 1997; vgl. dazu auch Peek, 2001).

Mit der Analyse von *Ergebnisrückmeldungen* an Schulen und Lehrkräfte beschäftigen sich neben unserem eigenen DFG-Projekt WALZER vor allem die

folgenden beiden Projekte, die wir aus Platzgründen allerdings nur nennen können, ohne ihre Ergebnisse zu berichten.

Projekt QuaSUM (**Qua**litätsuntersuchung an **S**chulen zum **U**nterricht in **M**athematik) (Lehmann, Gänsfuss & Peek, 2000a; Lehmann et al., 2000b). Im Kontext dieser Evaluationsstudie (Mathematik in 6. und 9. Klassen) lag ein besonderer Schwerpunkt auf der Frage nach der Nutzung dieser Daten für die schulische Qualitätsentwicklung, der Peek (1999; 2000) in einer Anschlussstudie nachging. Im Projekt QuaSUM wurde unseres Wissens erstmals systematisch die Rezeption von Ergebnisrückmeldungen untersucht.

Projekt QUASSU (Entwicklung und Implementation eines extern unterstützten Systems zur **Qua**litäts**s**icherung an **S**chulen unter besonderer Berücksichtigung des mathematisch-naturwissenschaftlichen **U**nterrichts): Dieses von Ditton geleitete und an frühere Untersuchungen des Autors anknüpfende Projekt hat in ausführlichen Zusatzuntersuchungen ebenfalls die Frage nach der Rezeption von Ergebnissen schulischer Leistungstests wie auch von Schülerwahrnehmungen des Unterrichts untersucht (Ditton, Arnoldt & Bornemann, 2002b). Erste Ergebnisse zur Ergebnisrückmeldung werden darüber hinaus von Ditton und Merz (2000) sowie von Ditton, Arnoldt, Babic, Bornemann & Zehme (2002a) berichtet.

MARKUS – die Basis für die Rezeptionsstudie WALZER

Anlage der Studie

Bei **MARKUS** (**Ma**thematik-Gesamterhebung **R**heinland-Pfalz: **K**ompetenzen, **U**nterricht, **S**chulkontext) handelt es sich um ein Evaluationsprojekt, das in allen Klassen der 8. Klassenstufe im Bundesland Rheinland-Pfalz (Bundesrepublik Deutschland) im Fach Mathematik durchgeführt wurde. Alle Schularten ausser Sonderschulen waren eingeschlossen. Die Mathematiktestleistungen der Schülerinnen und Schüler wurden mit einem eigens entwickelten, curricular validen Test erfasst. Zusätzlich wurden lern- und unterrichtsrelevante Bedingungsfaktoren auf Seiten der Schülerinnen und Schüler mittels eines Fragebogen erhoben. Ergänzend dazu wurden auch die Mathematiklehrkräfte der untersuchten Klassen sowie die Schulleiterinnen und Schulleiter mittels gesonderter Fragebögen befragt. Die Erhebungen fanden alle zum gleichen Zeitpunkt statt, nämlich am 31. Mai 2000. An der Untersuchung nahmen 37'520 Schülerinnen und Schüler in 1'876 Klassen an 625 Schulen teil. Details zur Anlage, Logistik und zu den Ergebnissen finden sich in Helmke & Jäger (2002). Im Folgenden beschränken wir uns auf die den Schulen zurückgemeldeten Ergebnisse, weil diese Ergebnisrückmeldung den Gegenstand der Rezeptionsstudie WALZER darstellt.

Ergebnisrückmeldungen

MARKUS war deswegen als flächendeckende Studie angelegt worden, weil über allgemeine Aussagen zum Leistungsniveau von Schülerinnen und Schülern innerhalb von Rheinland-Pfalz (system-monitoring) hinaus allen Schulen und Klassen eine *Rückmeldung* über ihren Leistungsstand und die Ausprägung lernrelevanter Merkmale gegeben werden sollte, um damit Prozesse des schulischen Qualitätsmanagements anzustossen. Für die Rückmeldungen waren zwei separate Schritte vorgesehen:

(1) Mitgeteilt wurden pro Klasse ein *Leistungsprofil* (Ergebnis der Klasse im *Gesamttest*, sowohl als Rohwert in der Metrik der Rasch-Skala wie auch als um Kontextunterschiede bereinigter Wert; von der Klasse in den einzelnen *Subtests* erreichte Werte) sowie ein *Kontextprofil* (Ausprägungen relevanter Kontextmerkmale). Die Rückmeldung erfolgte in Form von Prozentrangzonen, bei der die ursprüngliche Prozentrangskala in Intervalle von 1 bis 10 unterteilt worden war. Auf diese Weise war die Lage der Klasse im Vergleich zu allen anderen Klassen des gleichen Bildungsgangs ersichtlich. Diese Rückmeldung ging an alle Schulen, deren Schulleiter und die beteiligten Mathematiklehrkräfte sowie die Schulaufsichten. Etwa ein halbes Jahr später erfolgte eine zweite Rückmeldung, die noch weitergehende Erläuterungen umfasste (vgl. dazu die Musterrückmeldung in Helmke & Jäger, 2002, Anhang 4). Kern dieser Rückmeldung war das im Anhang Abb. 1 dargestellte, mit ausführlichen Erläuterungen versehene Leistungsprofil.

Als zusätzliche Einordnungshilfe enthielt die Leistungsrückmeldung ein *Schaubild* (vgl. Anhang Abb. 2) mit einer der Grösse nach geordneten Verteilung der Testleistungen (Gesamttestleistung) im jeweiligen Bildungsgang (Hauptschule Grundkursniveau, Hauptschule A-Kursniveau, Realschule, Gymnasium). Diese Darstellung sollte es den Lehrkräften auf eine anschauliche und leicht nachvollziehbare Weise ermöglichen, die von der eigenen Klasse erzielte Leistung mit den Leistungen aller anderen Klassen zu vergleichen.

Bestandteil der Leistungsrückmeldung war auch eine Rückmeldung zum Kontext, die den Lehrkräften Hinweise über lernerleichternde und lernerschwerende Bedingungen des Klassenkontextes geben sollte, die bei einer sachgemässen Beurteilung des Leistungsstandes der Klasse in Rechnung gestellt werden müssen (vgl. Anhang Abb. 3). Für detaillierte Angaben und ausführliche Erläuterungen sei auf die 16 Seiten umfassende Musterrückmeldung in Helmke & Jäger (2002, Anhang 4) verwiesen.

(2) Darüber hinaus konnten alle beteiligten Mathematiklehrkräfte ein *Profil* von *Unterrichts- und Lernermerkmalen* ihrer Klasse anfordern (vgl. die Musterrückmeldung in Helmke & Jäger, 2002, Anhang 3). Grundlagen dafür waren klassenweise aggregierte Angaben im Schülerfragebogen (Unterrichtsperzeption, motivationale Merkmale der Schülerinnen und Schüler). Über 60% aller Lehrkräfte machten von diesem Rückmeldungsangebot Gebrauch. Diese Rückmeldungen wurden an die von den Lehrkräften angegebene Adresse (je nach

Wunsch Privat- oder Schulanschrift) verschickt. Ob und in welchem Masse die Lehrkräfte diese Rückmeldungen in ihrer Schule bekannt gaben, blieb ihnen freigestellt.

Die Rezeptionsstudie WALZER

Ziel des Projekts WALZER war es zu analysieren, wie die Rückmeldungen rezipiert wurden und ob und welche Wirkungen sie in den Schulen entfalten. Das Projekt WALZER bot die einzigartige Möglichkeit, an eine gross angelegte und für die Schulen sehr bedeutsame Evaluationsstudie anzuknüpfen und deren Effekt zu überprüfen. Grundlage der geplanten Untersuchungen war ein umfassendes Modell der verschiedenen Faktoren, die für eine Rezeption der Rückmeldungen und deren Umsetzung in geeignete pädagogische Massnahmen relevant sind (vgl. Helmke & Schrader, 2001). Dazu fanden Befragungen der Lehrkräfte und der Schulleitungen statt.

Der vorliegende Beitrag beschränkt sich auf die Ergebnisse der *Schulleiterbefragung*[2] und damit auf die Schulebene, die der logische Ausgangspunkt für detailliertere Analysen ist. Die Überlegungen, die den Fragestellungen dieses Beitrags zugrunde liegen, lassen sich anhand des in Abbildung 2 dargestellten Modells verdeutlichen.

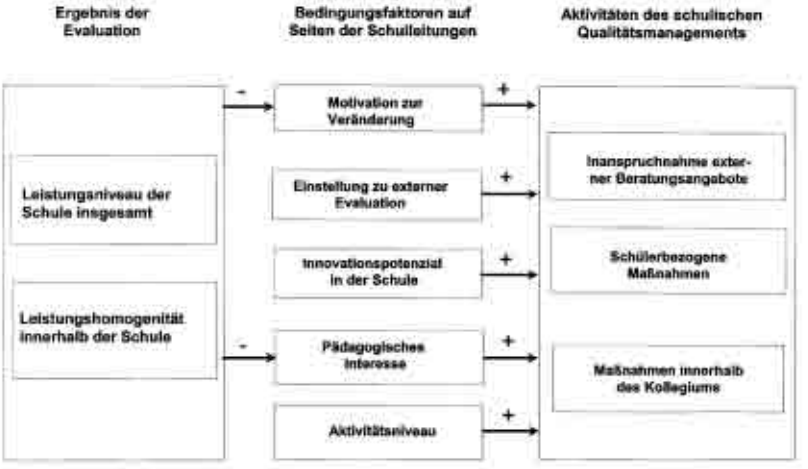

Abbildung 2: Annahmen zu Bedingungen schulischen Qualitätsmanagements auf der Schulebene

Im Mittelpunkt steht die Frage, ob die Evaluation Aktivitäten des schulischen Qualitätsmanagements bewirkt oder forciert hat (vgl. rechte Seite von Abbildung 2). Ausgangspunkt solcher Aktivitäten ist das im Leistungstest von MAR-

KUS erzielte Ergebnis (linke Seite von Abbildung 2). Ob und inwieweit das Leistungsergebnis einer Schule zu Aktivitäten und Massnahmen innerhalb der Schule führt, hängt von verschiedenen Bedingungsfaktoren ab. Änderungen innerhalb der Schule sind vor allem dann zu erwarten, wenn die Ergebnisse erwartungswidrig und nicht zufriedenstellend ausfallen (also ein niedriges *Leistungsniveau* der Schule und eine geringe innerschulische *Leistungshomogenität* anzeigen) und wenn den Ergebnissen eine hohe Relevanz für die Schule zuerkannt wird; beides sollte in einer hohen *Motivation zur Veränderung* resultieren. Mit der Einleitung von Massnahmen ist insbesondere dann zu rechnen, wenn bei Schulleitern und Lehrkräften eine *positive Einstellung zu externer Evaluation*, ein hohes *Innovationspotenzial* und ein *inhaltliches Interesse an pädagogischen Fragen* vorhanden sind, wenn der Schulleiter beim innerschulischen Transfer «Leadership» zeigt, d.h. eine aktive Rolle einnimmt und in breitem Umfang über die Ergebnisse der Evaluation informiert *(Aktivitätsniveau).* Im Einzelnen geht es um folgende Fragestellungen:

- Wie sind die für Aktivitäten des schulischen Qualitätsmanagements massgeblichen Bedingungsfaktoren (Einschätzungen und Einstellungen der Schulleiterinnen und Schulleiter) beschaffen?
- Führte MARKUS zur Inanspruchnahme externer Beratungsangebote und zur Initiierung (oder zum Ausbau) innerschulischer Massnahmen der Qualitätssicherung und Schulentwicklung,
- und hängen diese Aktivitäten in erwarteter Weise mit den Bedingungsfaktoren auf Seiten der Schulleitungen zusammen?

Obwohl Abbildung 2 als Analysestrategie ein Strukturgleichungsmodell (wie LISREL, AMOS, EQS) nahe legt, belassen wir es im Folgenden bei einfachen korrelativen Analysen, da die Voraussetzungen für komplexe Pfadmodelle (Stichprobengrösse, Mehrfachverankerung der Konstrukte) nicht gegeben sind.

Methode

Stichprobe

Die Analysen basieren auf einer Stichprobe von N = 52 Schulleitern. 11,5% sind weiblich, 88,5% männlich. Das mittlere Alter liegt bei knapp 54 Jahren; die Altersbereiche sind wie folgt vertreten: 36 bis 40 Jahre mit 1,9%; 41 bis 45 Jahre mit 3,8%; 46 bis 50 Jahre mit 23,1%; 51 bis 55 Jahre mit 34.6%; 56 bis 60 Jahre mit 23,1%; über 60 Jahre mit 13,5%. Von den befragten Schulleiterinnen und Schulleitern unterrichten 48,1% selbst das Fach Mathematik. Die Fragebogen waren an die Leitungen aller beteiligten 625 Schulen verschickt worden, so dass der Rücklauf als sehr gering bezeichnet werden muss. Auf die damit verbundenen Probleme wird an späterer Stelle noch genauer eingegangen.

Instrumente

Der Schulleiterfragebogen umfasste schwerpunktmässig Angaben zur Schule und zur Person, zu den Leistungsrückmeldungen und zu Massnahmen der Qualitätssicherung und Schulentwicklung. Die deskriptiven Statistiken der verwendeten Variablen einschliesslich der Reliabilitäten (sofern es sich um Skalen handelt) sind in Tabelle 1 dargestellt.

Tabelle 1: Deskriptive Statistiken der verwendeten Variablen

Variable	Itemzahl	Wertebereich	M	SD	Schiefe	α
Leistungsniveau der Schule	max. 10	1 - 10	5.52	2.01	0.15	-
Innerschulische Leistungshomogenität	max. 10	≥ 0	1.97	0.77	-0.15	-
Motivation zur Veränderung	4	1 – 4	2.45	0.59	0.33	0.81
Einstellung zu externer Evaluation	8	1 - 4	2.70	0.55	-0.29	0.87
Innovationspotenzial innerhalb der Schule	11	1 - 4	3.26	0.38	-.26	0.83
Pädagogisches Interesse	11	1 - 3	2.11	0.30	-0.56	0.86
Aktivitätsniveau	8	0 - 1	0.32	0.24	0.29	0.61
Schulisches Qualitätsmanagement	16	0 - 2	0.19	0.21	0.98	0.71
Inanspruchnahme von externer Beratung	4	0 - 4	0.67	0.83	1.72	0.79

Leistungsniveau der Schule und innerschulische Leistungshomogenität. Es wurde gefragt, welche Prozentrangzone (von 1 = niedrig bis 10 = hoch) die Klassen bzw. Kurse seinerzeit in der MARKUS-Untersuchung erreicht hatten. Der Mittelwert dieser Angaben wird als Indikator für das Leistungsniveau der Schule und die schulinterne Streuung dieser Angaben als Mass für die innerschulische Leistungshomogenität herangezogen.

Motivation zur Veränderung. Diese Skala basiert auf vier Items (Erwartungswidrigkeit der Ergebnisse, Zufriedenheit mit den Ergebnissen, subjektive Relevanz der Ergebnisse und Einschätzung der Notwendigkeit von Veränderungen).

Einstellung zu externer Evaluation. Die 8 Items (7 davon wurden von Ditton & Merz, 2000 übernommen) erfassen, wie die Befragten die Bedeutung der Evaluation einschätzen. Itembeispiel: «Gross angelegte Evaluationsstudien, wie z.B. TIMSS, LAU, QuaSUM, MARKUS liefern wichtiges Steuerungswissen für das Bildungswesen». Antwortmöglichkeiten: stimme voll zu/stimme eher zu/stimme eher nicht zu/stimme gar nicht zu.

Innovationspotenzial in der Schule. Für die Erfassung der Innovationsbereitschaft wurden 11 Items verwendet. Ausgangspunkt waren einzelne Items des Schulbarometers (Institut für Schulentwicklungsforschung, 1999), die erheblich modifiziert und erweitert wurden. Itembeispiel: «In unserer Schule spielen Fragen der Entwicklung und Erneuerung eine grosse Rolle». Antwortmöglichkeiten: stimme voll zu/stimme eher zu/stimme eher nicht zu/stimme gar nicht zu.

Pädagogisches Interesse. Hierbei geht es speziell um das Interesse an Faktoren, die für den Lern- und Unterrichtserfolg bedeutsam sind. Die 11 selbst entwickelten Items liessen sich zu einer reliablen Skala zusammenfassen. Itembeispiel: «Nachfolgende Themen sollten im Abschlussbericht behandelt werden: Rolle der Klassengrösse». Antwortmöglichkeiten: intensiv/begrenzt/gar nicht. Klassengrösse, Lehrerbelastung und Unterrichtsausfall gehen dabei mit umgekehrter Polung in die Skala ein.

Aktivitätsniveau. In dieser 8-Item-Skala ging es darum, ob und wie intensiv sich die Schulleitung für die Besprechung von Rückmeldungen engagierte, z.B. ob die Rückmeldungen einzeln mit den Lehrkräften besprochen wurden.

Schulisches Qualitätsmanagement. Erfasst wurden drei verschiedene Facetten: (1) *Massnahmen innerhalb des Kollegiums* (8 Items). Itembeispiel: «Arbeitsgruppen zu neuen Aspekten der Fachdidaktik»; (2) *schülerbezogene Massnahmen* (8 Items). Itembeispiel: «Ergänzende Kurse in Mathematik für besonders leistungsstarke Schüler/innen». Antwortmöglichkeiten: infolge von MARKUS initiiert/infolge von MARKUS ausgebaut; die Skalierung erfolgte so, dass «ausgebaut» mit 1 und «initiiert» mit 2 kodiert wurde; (3) *Inanspruchnahme von externer Beratung* (4 Items). Itembeispiel: «Welche Beratungsangebote wurden an Ihrer Schule im Zusammenhang mit MARKUS in Anspruch genommen? Beratung durch das IFB oder andere Service-Einrichtungen». Antwortmöglichkeiten: nein/ja. Einzelne Items wurden aus der MARKUS-Untersuchung (Helmke & Jäger, 2002) übernommen. Für die ersten beiden Facetten liessen sich keine separaten Skalen bilden; die Skala «schulisches Qualitätsmanagement» basiert auf Items aus beiden Bereichen.

Ergebnisse

Repräsentativität und deskriptive Ergebnisse

Angesichts des geringen Rücklaufs ist die Frage von besonderer Bedeutung, wie repräsentativ die befragten Schulleiterinnen und Schulleiter sind. Hinweise dazu ergeben sich zunächst aus dem Vergleich mit der Gesamtpopulation der Schulleiter/innen, die die Zielgruppe der Schulleiterbefragung innerhalb der MARKUS-Studie bildete. Da aus Gründen des Datenschutzes und der Akzeptanz eine direkte Zuordnung der in WALZER gewonnenen Befragungsergebnisse zu den

MARKUS-Daten nicht möglich war, können nur prozentuale Häufigkeiten verglichen werden.

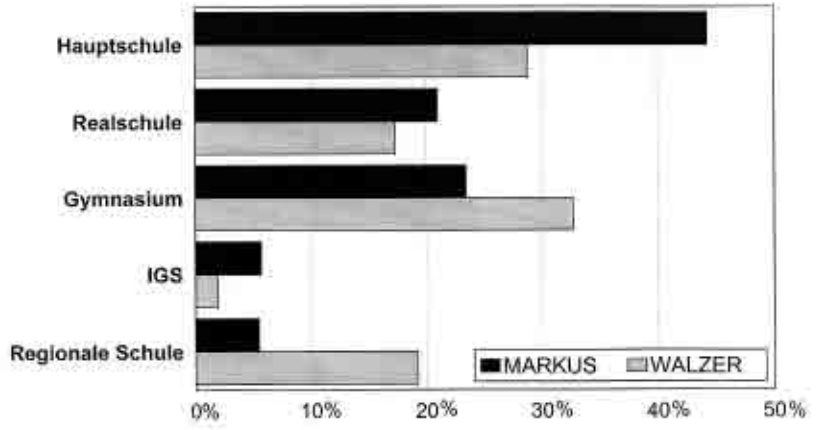

Abbildung 3: Vergleich der Schularten in MARKUS und WALZER

Ein Vergleich der *Schularten* (vgl. Abbildung 3) zeigt, dass die Hauptschulen in der WALZER-Schulleiterbefragung unter- und die Gymnasien überrepräsentiert sind. Deutlich überrepräsentiert ist vor allem die relativ kleine Gruppe der Regionalen Schulen. Auch wenn das Verteilungsmuster der WALZER-Stichprobe nicht allzu stark von der bei MARKUS gefundenen Verteilung abzuweichen scheint, sind die Abweichungen hochsignifikant, (χ^2 (4, N = 52) = 24,38, p< 0.001.

Sind die untersuchten Klassen und Schulen im Hinblick auf ihren *Leistungsstand* repräsentativ? Der Mittelwert des erfragten schulischen Leistungsniveaus (d.h. der an der Schule im Durchschnitt errreichten Prozentrangzonen) weicht lediglich um 0.02 von dem theoretisch zu erwartenden Gesamtmittelwert von 5,5 ab, t (51) = 0.05, n.s. (bezogen auf die für die einzelnen Schularten zu erwartenden Werte ergeben sich folgende Abweichungen: 0,09 bei den Hauptschulen, -0,19 bei den Realschulen, 0,03 bei den Gymnasien, -0,40 bei den Regionalen Schulen und -1,77 bei den Intergrierten Gesamtschulen (IGS); kein Unterschied ist signfikant). Bis auf die Integrierten Gesamtschulen, die nur mit einer einzigen (deutlich unter dem Leistungsdurchschnitt liegenden) Schule in der Stichprobe vertreten sind, weichen also die schulartspezifischen Mittelwerte nur geringfügig vom Gesamtmittelwert ab. D.h. die untersuchten Schulen und Klassen können hinsichtlich ihres Leistungsstandes als repräsentativ für die Gesamtheit der untersuchten Schulen und Klassen angesehen werden.

In Tabelle 2 wird die Einstellung der Schulleiterinnen und Schulleiter zur externen Evaluation dargestellt.

Tabelle 2: Einstellung zu externer Evaluation

	Antwortverteilung (in Prozent)					
Gross angelegte Evaluationsstudien, wie z.B. TIMSS, LAU, QuaSUM, MARKUS ...	stimme gar nicht zu	stimme eher nicht zu	stimme eher zu	stimme voll zu	M	M_V
liefern wichtiges Steuerungswissen für das Bildungswesen	2.0	32.0	52.0	14.0	2.78	-
sind für meine Arbeit als Schulleiter/in nützlich	4.1	22.5	67.3	6.1	2.75	-
sind für die Arbeit der Schulen sehr wichtig	6.0	26.0	56.0	12.0	2.74*	3.02
tragen dazu bei, dass man sich in den Schulen mehr bemüht	11.8	33.3	49.0	5.9	2.49**	2.86
geben eine objektive Basis ab, um zu sehen, wo eine Schule steht.	18.4	34.7	40.8	6.1	2.35**	2.83
nützen für die eigentliche Arbeit der Lehrer/innen wenig (u)	14.0	40.0	42.0	4.0	2.36(*)	2.16
bringen nur Unruhe in die Schulen (u)	22.4	42.9	32.7	2.0	2.14*	1.88
schaffen mehr Probleme als sie nützen (u)	26.0	48.0	24.0	2.0	2.02*	1.76

Anmerkung. u: Diese Items werden bei der Skalenbildung umgepolt. M: mittlere Einschätzung (stimme gar nicht zu = 1, stimme eher nicht zu = 2, stimme zu = 3, stimme voll zu = 4); M_V: Vergleichswert bei Ditton, Merz und Edelhäusser (2002c); (*) $p < 0.10$; * $p < 0.05$; ** $p < 0.01$.

Die Items in Tabelle 2 sind nach dem Ausmass der Zustimmung zu der jeweiligen Aussage geordnet; d.h. die Items mit höherer Zustimmung stehen in der Tabelle oben. Die für die spätere Skalenbildung umzupolenden Items sind unten separat aufgeführt. Es zeigt sich, dass vor allem Items günstig eingeschätzt werden, die einen eher allgemein gehaltenen Nutzen ansprechen. Vergleicht man diese Ergebnisse mit denen einer grösseren Zufallsstichprobe von Schulleitern (N = 169) in der Untersuchung von Ditton, Merz und Edelhäusser (2002c), so finden sich bei den Teilnehmern der vorliegenden Untersuchung deutlich ungünstigere Einstellungen zur Evaluation (vgl. Tabelle 2). Dies spricht gegen die Annahme, es handele sich bei den Befragten überwiegend um Schulleiter/innen, die der Evaluation äusserst positiv gegenüber stehen.

Angaben zum Innovationspotenzial in der Schule aus Sicht der Schulleiter/innen sind in Tabelle 3 dargestellt. Hier zeigt sich, dass das Veränderungspotenzial der eigenen Schule relativ hoch eingeschätzt wird. Am günstigsten fallen Einschätzungen aus, die die eigene Rolle und solche Aspekte der eigenen Schule betreffen, die im Verantwortungs- und Einflussbereich des Schulleiters liegen. Auch hier sind die für die Bildung der Skala umzupolenden Items separat dargestellt.

Tabelle 3: Innovationspotential einer Schule aus Sicht der Schulleiter/innen

	Antwortverteilung (in Prozent)			
Wie stehen Sie zu den folgenden Aussagen, die sich auf Veränderungen in der Schule beziehen?	stimme gar nicht zu	stimme eher nicht zu	stimme eher zu	stimme voll zu
Veränderungen, die pädagogisch sinnvoll sind, werden von mir nachhaltig unterstützt	0.0	0.0	23.1	76.9
An unserer Schule werden neue Unterrichtsformen und -methoden unterstützt	0.0	5.8	44.2	50.0
Ich versuche selbst sehr stark, Innovationsprozesse in Gang zu setzen	0.0	5.8	50.0	44.2
Unsere Schule ist aufgeschlossen gegenüber neuen Entwicklungen	0.0	9.6	50.0	40.4
Ideen und Verbesserungsvorschläge werden an unserer Schule ernst genommen	0.0	5.8	61.5	32.7
In unserer Schule spielen Fragen der Entwickung und Erneuerung eine grosse Rolle	0.0	5.0	29.0	18.0
Ich versuche, Innovationen an unserer Schule ganz systematisch zu planen	1.9	13.5	53.8	30.8
Die Lehrkräfte an unserer Schule brauchen mehr Kenntnisse über alternative Unterrichtsformemen und -methoden	1.9	23.1	44.2	30.8
In unserer Schule bemüht man sich um eine systematische Weiterbildung	0.0	21.2	53.8	25.0
Vor Veränderungen wird stets ein grundlegendes Einvernehmen im Kollegium hergestellt	0.0	2.0	49.0	49.0
Innovationen werden bei uns planvoll und systematisch realisiert	0.0	25.0	59.6	15.4
Die Lehrkräfte unserer Schule sind von sich aus sehr an Innovationen interessiert	0.0	31.4	51.0	17.6
Die Arbeitsbelastung der Lehrkräfte lässt keinen Spielraum für die Erprobung neuer Maßnahmen (u)	2.0	37.2	49.0	11.8
Neue Vorgehensweisen sind selten besser als die, die sich bereits bewährt haben (u)	9.8	66.7	23.5	0.0

Anmerkung. u: Diese Items werden bei der Skalenbildung umgepolt.

Die in Tabelle 1 dargestellten deskriptiven Statistiken der verwendeten Variablen zeigen, dass bis auf «Aktivitätsniveau» alle verwendeten Skalen ausreichend reliabel sind. Auch bei den Leistungswerten der Schulen ergeben sich hinreichend grosse Streuungen.

Innerschulischer Umgang mit den MARKUS-Rückmeldungen

Wie werden die erzielten Ergebnisse von den Schulleitungen eingeschätzt? Im Folgenden werden zunächst die Ergebnisse zu den vier Items dargestellt, die in die Skala *Motivation zur Veränderung* eingegangen sind. Wie aus Abbildung 4

hervorgeht, fielen die Ergebnisse für fast 60% der Befragten erwartungsgemäss aus. Die Anteile derjenigen, deren Erwartungen übertroffen oder enttäuscht wurden, halten sich etwa die Waage.

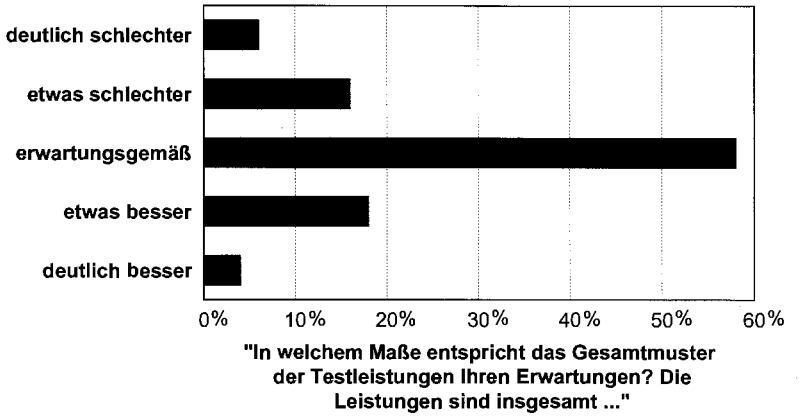

Abbildung 4: Einschätzung des MARKUS-Ergebnis durch die Schulleiter/innen

Der Anteil der Schulleiter/innen, die mit dem Ergebnis zufrieden sind, ist geringfügig höher als der Anteil der Unzufriedenen (vgl. Abbildung 5).

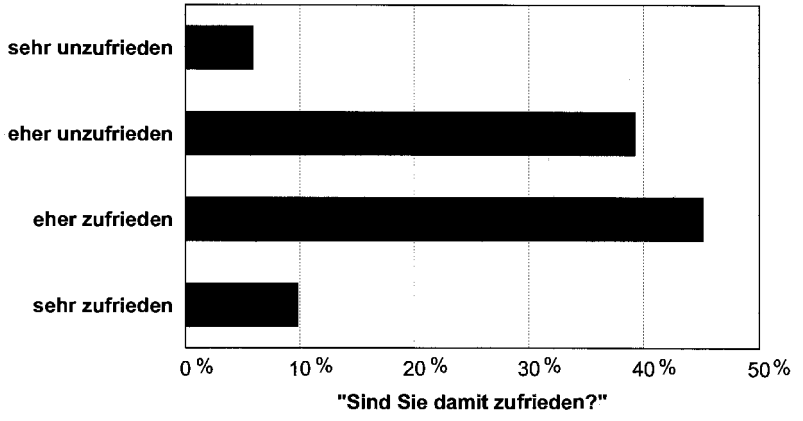

Abbildung 5: Zufriedenheit der Schulleiter/innen mit MARKUS-Ergebnis

Wie aus Abbildung 6 hervorgeht, misst nur ein sehr kleiner Teil der Befragten den Ergebnissen *überhaupt keine* Bedeutung zu. Für den grössten Teil der Schulleiter/innen haben die Ergebnisse zumindest *eine geringe* Bedeutung; und

immerhin ein Viertel von ihnen hält die Ergebnisse für *ziemlich* bis *sehr bedeutsam* für das schulische Qualitätsmanagement.

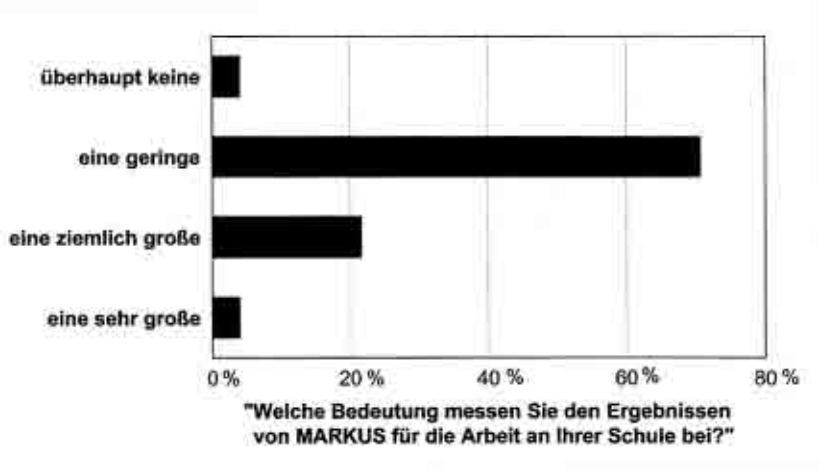

Abbildung 6: Bedeutung des MARKUS-Ergebnisses aus Sicht der Schulleiter/innen

Noch grösser ist der Anteil der Befragten, die aufgrund der MARKUS-Ergebnisse Anlass sehen, in der Schule etwas zu ändern (vgl. Abbildung 7). Möglicherweise geht es den Schulleiterinnen und Schulleiter bei dieser Frage stärker um grundsätzliche Veränderungen in der Schule, während die vorangegangene Frage eher auf die tagtägliche Unterrichtsarbeit bezogen wird.

Abbildung 7: Notwendigkeit von Änderungen aufgrund der MARKUS-Ergebnisse aus Sicht der Schulleiter/innen

An welchen pädagogischen Themen sind Schulleiterinnen und Schulleiter am stärksten interessiert? Die Ergebnisse dazu sind Abbildung 8 zu entnehmen.

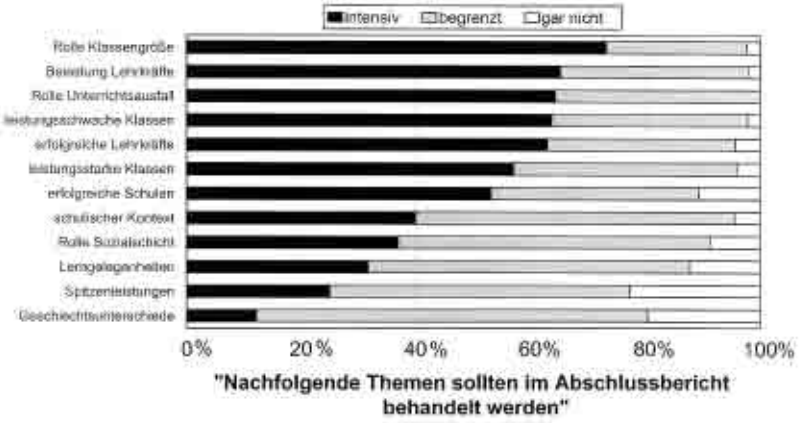

Abbildung 8: Pädagogisches Interesse

Spitzenreiter sind Themen, die grundlegende Rahmenbedingungen des Unterrichts ansprechen, die die Tätigkeit von Lehrkräften unmittelbar berühren und auch in der Öffentlichkeit intensiv diskutiert werden.

Wie die Schulleiterinnen und Schulleiter mit den erhaltenen Rückmeldungen umgehen, ist in Abbildung 9 dargestellt.

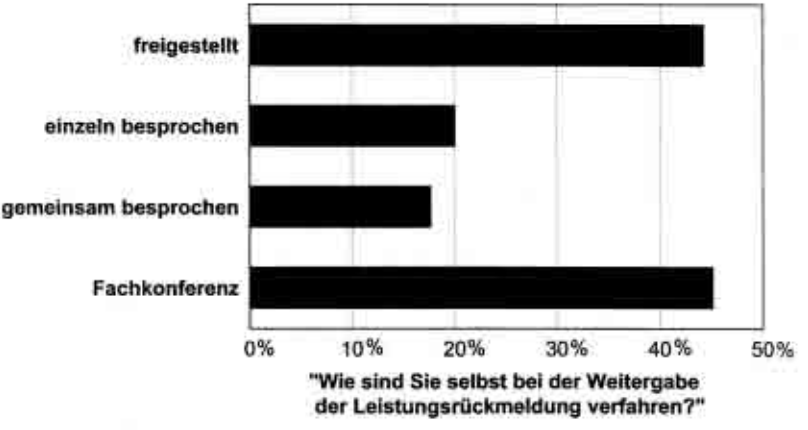

Abbildung 9: Umgang der Schulleiter/innen mit den Leistungsrückmeldungen (Aktivität beim innerschulischen Transfer)

Zwei Vorgehensweisen stehen im Vordergrund: Entweder wird den Lehrkräften freigestellt, ob und wie sie sich mit den Rückmeldungen auseinandersetzen, oder

die Rückmeldungen werden zum Gegenstand einer Fachkonferenz gemacht. In selteneren Fällen werden die Rückmeldungen einzeln oder gemeinsam mit den betroffenen Lehrkräften besprochen. Diese Angaben sind Bestandteil der Skala Aktivität beim innerschulischen Transfer, in die das erste Item («Lehrkräften wird die Ergebnisnutzung freigestellt») mit negativem Gewicht eingeht.

Massnahmen des schulischen Qualitätsmanagements

Wurden aufgrund der MARKUS-Ergebnisse Massnahmen der Qualitätssicherung in die Wege geleitet? Dies ist eine Frage, die an den Kern der mit diesen Evaluationsstudien verbundenen Hoffnungen und Befürchtungen rührt. Zunächst zur Inanspruchnahme *externer Beratungsangebote*: Am häufigsten (46,3%) wurde um eine Beratung durch Mathematikmoderatoren nachgesucht; mit Abstand folgt die Beratung durch Fortbildungs- und Serviceinstitutionen (14,3%). Gut 48% aller Schulen haben überhaupt kein Beratungsangebot in Anspruch genommen, 40% ein einziges Angebot, knapp 9% der Schulen zwei und gut 2% vier Angebote.

Massnahmen innerhalb des Kollegiums und schülerbezogene Massnahmen, die infolge der MARKUS-Studie initiiert oder ausgebaut wurden, sind in Abbildung 10 dargestellt. Hier geht es um das, was die Rückmeldung aus Sicht der Schulleiter/innen substantiell bewirkt hat.

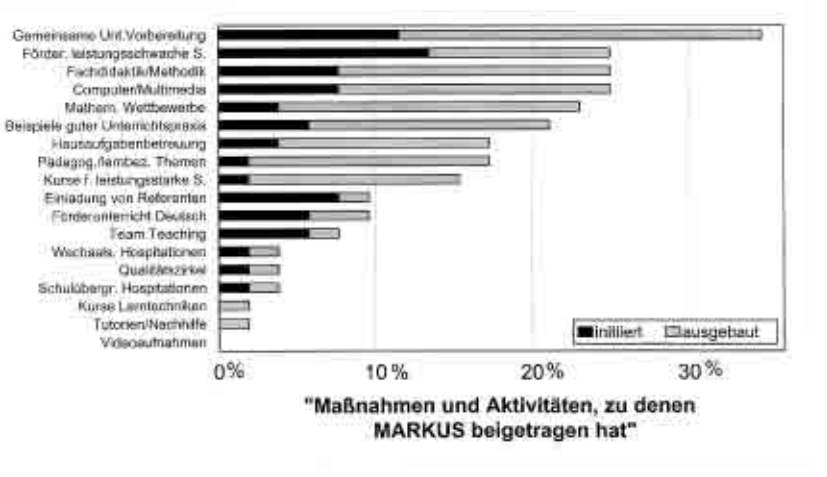

Abbildung 10: Als Folge von MARKUS eingeleitete Massnahmen des schulischen Qualitätsmanagements

Dabei gibt es fünf Spitzenreiter, die alle einen relativ engen Fachbezug aufweisen. Die *gemeinsame Vorbereitung des Mathematikunterrichts* als das bei weitem am häufigsten genannte Merkmal wird möglicherweise als ein effektiver, mit vergleichsweise wenig Aufwand und grossem potentiellen Nutzen verbundener Ein-

stieg in kooperative Arbeitsstrukturen gesehen. Bereits bei MARKUS hatte sich gezeigt, dass die gemeinsame Unterrichtsvorbereitung bei einem nicht unbeträchtlichen Prozentsatz aller Klassen vorkommt (in der Hauptschule immerhin bei 25% aller Lehrkräfte, im Hauptschulzweig der integrierten Gesamtschulen bei über 45% aller Lehrkräfte; Helmke, Hosenfeld, Schrader & Wagner, 2002). Die Bildung von *Arbeitsgruppen* zu neuen Aspekten der *Fachdidaktik und Methodik* geht vermutlich in die gleiche Richtung eines zunehmenden Austausches innerhalb des Kollegiums. Vielleicht haben die seit TIMSS oft als Vorbild hingestellten Verhältnisse an japanischen Schulen bereits Anklang gefunden. Im oberen Mittelfeld findet sich ein recht heterogenes Bündel von Massnahmen, die von der Erarbeitung und Bereitstellung schulinterner Beispiele guter Unterrichtspraxis bis zur Durchführung von Kursen für leistungsstarke Schüler reichen. Das untere Mittelfeld umfasst Massnahmen wie die Einladung von Referenten für die schulinterne Lehrerfortbildung bis hin zur gemeinsamen Unterrichtsdurchführung (Team Teaching). Dass Team Teaching überaus selten ist (Ausnahme ist wieder der Hauptschulzweig der IGS), hatte sich ebenfalls in der MARKUS-Studie gezeigt (Helmke et al., 2002). Sehr viel seltener genannt werden Themen wie schulinterne und schulübergreifende Hospitationen und Qualitätszirkel. Dies könnte darauf hindeuten, dass sich Lehrkräfte wohl immer noch stark als Einzelkämpfer sehen und von einer sich entwickelnden Kultur der schulinternen Kooperation noch nicht die Rede sein kann. Unterrichtsergänzende Kurse in allgemeinen Lerntechniken sowie Tutorien bzw. Nachhilfe durch Lehrkräfte werden selten genannt. Ein vielversprechendes Mittel der Unterrichtsentwicklung, der Einsatz von Videoaufzeichnungen des Unterrichts (Helmke, 2003), wird in den untersuchten Schulen (zumindest nach Einschätzung der Schulleiter/innen) überhaupt nicht genannt.

In Abbildung 11 ist die Anzahl der pro Schule vorkommenden qualitätssichernden Massnahmen dargestellt (getrennt danach, ob es sich um den Ausbau bestehender oder die Initiierung neuer Massnahmen handelt). Hier zeigt sich, dass in einem beträchtlichen Anteil der Schulen überhaupt keine Massnahmen infolge von MARKUS ausgebaut oder initiiert wurde. Auf der anderen Seite gibt es Schulen, bei denen bis zu acht unterschiedliche Massnahmen ausgebaut bzw. bis zu sechs Massnahmen initiiert wurden.

Wichtig zur Vermeidung von Fehlinterpretationen: Die in Abbildung 11 dargestellte Verteilung gibt kein Gesamtbild der tatsächlich in den Schulen ablaufenden Aktivitäten. Gefragt wurde nur, ob die Massnahmen *infolge von MARKUS* ausgebaut oder initiiert wurden. Über Massnahmen des schulischen Qualitätsmanagements, die *nicht* auf MARKUS zurückführbar sind, sagt Abbildung 11 nichts aus.

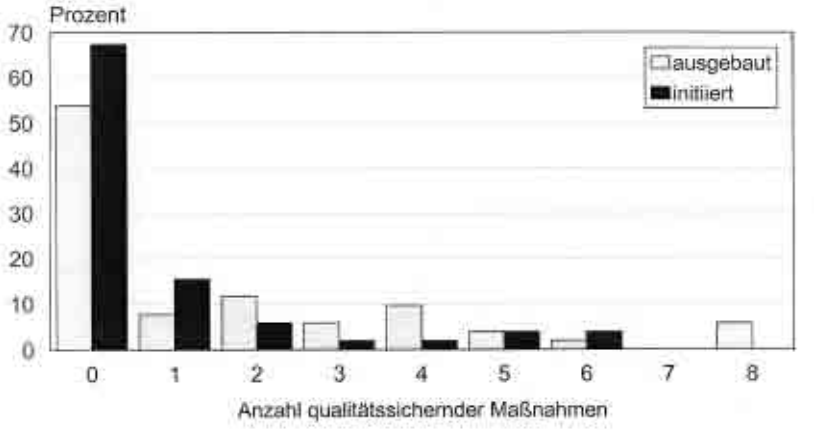

Abbildung 11: Verteilung der Anzahl qualitätssichernder Massnahmen

Zusammenhänge zwischen schulischem Qualitätsmanagement und ausgewählten Bedingungsfaktoren

Abschliessend sollen die aufgrund des in Abbildung 2 dargestellten Bedingungsmodells zu erwartenden Zusammenhänge überprüft werden. In Tabelle 4 sind zunächst die einfachen Korrelationen zwischen dem Leistungsergebnis aus der MARKUS-Studie und den Bedingungsfaktoren für das innerschulische Qualitätsmanagement dargestellt. Ein niedriges Leistungsniveau geht erwartungsgemäss mit einer hohen Motivation zur Veränderung einher; tendenziell gilt dies auch für eine grosse Leistungsheterogenität innerhalb der Schule. Dies dürfte Ausdruck des plausiblen Sachverhalts sein, dass ein ungünstiges Evaluationsergebnis einen gewissen Änderungsdruck erzeugt hat. Unsere Erwartung, dass ein ungünstiges Leistungsergebnis das pädagogische Interesse beinflusst, wird für das Leistungsniveau, nicht aber die Leistungshomogenität bestätigt. Die Ergebnisse zum Innovationspotenzial, für das keine Hypothese aufgestellt worden war, deuten darauf hin, dass auch dieses Merkmal vom Leistungsergebnis der Schule beeinflusst sein könnte und dass eine hohe Innovationsbereitschaft vor allem dann gegeben ist, wenn ein niedriges Leistungsniveau der Schule mit grosser Leistungshomogenität zusammentrifft (also alle Klassen gleichermassen schlechte Ergebnisse erzielt haben). Gruppiert man die Schulen nach hoher und niedriger Ausprägung beider Merkmale, so zeigt sich in der Tat, dass die Innovationsbereitschaft bei diesen Schulen am höchsten ist, ohne dass sich dies allerdings statistisch absichern liesse.

Tabelle 4: Zusammenhänge zwischen Evaluationsergebnis und Bedingungsfaktoren für ein schulisches Qualitätsmanagement

	Leistungsniveau der Schule	Leistungshomogenität innerhalb der Schule
Motivation zur Veränderung	-.61***	-.22 (*)
Einstellung zu externer Evaluation	-.19	-.12
Innovationspotenzial in der Schule	-.23 (*)	.28*
Pädagogisches Interesse	-.30*	-.09
Aktivitätsniveau	.05	-.19

Anmerkungen: (*) $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; einseitige Testung

Die in Tabelle 5 dargestellten Korrelationen zeigen, dass sich die Inanspruchnahme von Beratung (unterschieden werden hier die Ausprägungen «keine» «eine» und «mehr als eine» Beratung) nicht mit den im Modell zugrunde gelegten Bedingungsfaktoren in Zusammenhang bringen lässt. Bedeutsam für die Einleitung von Veränderungen ist weniger das Leistungsergebnis selbst, sondern die aktuelle Motivation zur Veränderung (der wahrgenommene Veränderungsdruck), die Einstellung zur externen Evaluation und das pädagogische Interesse, während das Innovationspotenzial ebenso wie die aktive Rolle des Schulleiters beim Umgang mit den Evaluationsergebnissen nicht zur Vorhersage beiträgt. Dies deutet darauf hin, dass ein bestimmtes Evaluationsergebnis für sich genommen nicht unbedingt zur Einleitung bestimmter Massnahmen führt, sondern möglicherweise erst dann, wenn bestimmte aktivitätsförderliche oder -erleichternde Einstellungen und Motivationslagen vorhanden sind bzw. aktiviert werden.

Tabelle 5: Zusammenhang zwischen Inanspruchnahme von Beratung, schulischem Qualitätsmanagement und ausgewählten Bedingungsfaktoren

	Inanspruchnahme von Beratung	Schulisches Qualitätsmanagement
Leistungsniveau der Schule	.18	-.19
Leistungshomogenität der Schule	-.02	-.23 (*)
Motivation zur Veränderung	-.06	.35*
Einstellung zu externer Evaluation	-.09	.45**
Innovationspotenzial in der Schule	.01	-.07
Pädagogisches Interesse	-.15	.24*
Aktivität beim innerschulischen Transfer	.12	.12

Anmerkungen: (*) $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; einseitige Testung

Statistisch signifikante Wechselwirkungen zwischen den Faktoren lassen sich allerdings nicht nachweisen. Ingesamt gesehen lassen sich durch die drei signifikanten motivationalen Merkmale 23,8% der Varianz des schulischen Qualitätsmanagements erklären. Wenn man diese Merkmale im Rahmen einer schritt-

weisen Regression simultan analysiert, dann leistet jedoch nur noch die Einstellung zur externen Evaluation einen signifikanten Erklärungsbeitrag.

Abschliessend sollen die Bedingungsprofile von zwei Extremgruppen verglichen werden: Es gibt eine relativ grosse Gruppe von Schulen (34.6%), die – so die Schulleiterbefragung – in der Folge von MARKUS überhaupt keine Initiativen zur Qualitätsverbesserung entwickelt hat. Stellt man dieser Extremgruppe diejenigen knapp 20% der Schulen gegenüber, die die meisten Anstrengungen im Bereich des schulischen Qualitätsmanagements unternommen haben, so ergibt sich folgendes Bild (vgl. Abbildung 12).

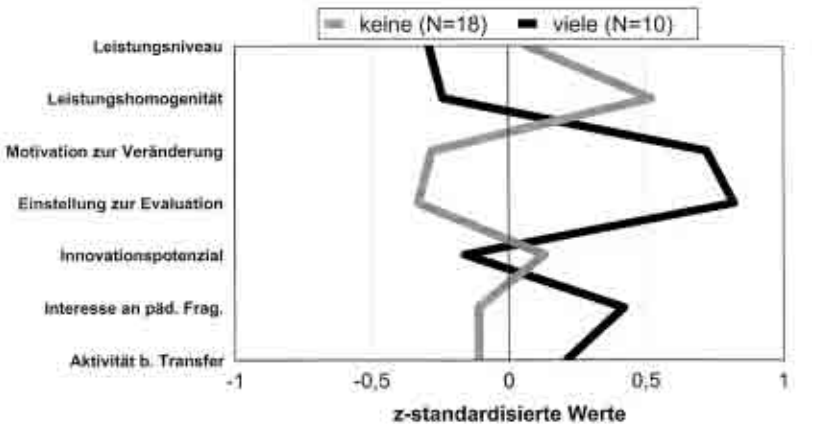

Abbildung 12: Vergleich der Schulen (Extremgruppen), die keine bzw. viele Massnahmen des schulischen Qualitätsmanagements ausgebaut bzw. eingeleitet haben

Auch hier wird deutlich, dass die Motivation zur Veränderung, $F(1,26) = 7,91$, $p < 0.01$, und die Einstellung zur Evaluation, $F(1,26) = 10.45$, $p < 0.01$, die grössten und – angesichts der jetzt stark reduzierten Stichprobe nicht verwunderlich – einzigen signifikanten Beiträge zur Unterscheidung der beiden Gruppen leisten. Bis auf das Innovationspotenzial, das überhaupt nicht zur Unterscheidung beiträgt, liegen immerhin alle Unterschiede in der erwarteteten Richtung.

Diskussion

Zunächst ist auf den geringen Rücklauf der Fragebögen und die daraus resultierende ungesicherte Repräsentativität einzugehen. Die untersuchten Schulen weichen hinsichtlich der Verteilung der Schularten, nicht aber beim Leistungsstand von der Gesamtheit der rheinland-pfälzischen Schulen ab. Differenziertere Ver-

gleichskriterien fehlen jedoch. Vergleicht man die Einstellungen der Untersuchungsteilnehmer zur Evaluation mit den Ergebnissen der Untersuchung von Ditton et al. (2002c), so sind in der eigenen Studie überraschenderweise sogar etwas ungünstigere Einschätzungen zu verzeichnen. Vor diesem Hintergrund erscheint es unwahrscheinlich, dass sich die Schulleiterinnen und Schulleiter, die den Fragebogen beantwortet haben, von denen, die nicht geantwortet haben, durch eine günstigere Einstellung zur Evaluation unterscheiden.

Wie ist die niedrige Beteiligung an der Befragung zu erklären? Neben der allgemeinen Belastung durch eine Vielzahl von Aufgaben und Verpflichtungen, denen Schulleiterinnen und Schulleiter ausgesetzt sind und die sie möglicherweise davon abgehalten haben, sich einer nicht besonders vordringlichen und folgenreichen Aktivität wie der Bearbeitung eines Fragebogens zu widmen, sehen wir dafür im Wesentlichen drei Gründe:

Timing. Aus Sicht der Beteiligten ist die Klage verständlich, dass die Rückmeldungen erst mit einer Verzögerung von einem halben Jahr verfügbar waren. Da die Evaluationsuntersuchung gegen Ende des Schuljahres stattfand, konnten die Ergebnisse für die pädagogische Arbeit in den untersuchten Klassen nur dann direkt genutzt (z.B. durch gezielte Förderung defizitärer Bereiche) werden, wenn die Mathematiklehrkraft die Klasse auch noch im folgenden Schuljahr unterrichtete. Die erhebliche zeitliche Verzögerung – bedingt durch die umfänglichen Prozesse des Einlesens, der Prüfung und der Analyse der Daten – stellt ein allgemeines Dilemma solcher grossangelegten Untersuchungen dar, das keinesfalls nur die vorliegende Untersuchung betrifft; so war bei PISA der zeitliche Abstand zwischen Erhebung und Schulrückmeldung mehr als doppelt so gross wie hier.

Sättigungseffekt. Zum Zeitpunkt der WALZER-Erhebung hatte sich das Interesse an den rückgemeldeten Ergebnissen stark reduziert, nicht zuletzt auch unter dem Einfluss der aufkommenden Diskussion um die PISA-Ergebnisse. Dies mag ebenso wie die in vielen Schulen des Landes verbreitete Abneigung gegen eine «von oben verordnete» Evaluation dazu beigetragen haben, dass Lehrkräfte und Schulleiter/innen nicht mehr bereit waren, an einer weiteren Befragung teilzunehmen. Möglicherweise kann der geringe Rücklauf der Fragebögen trotz grundsätzlich positiver Einstellung zur Evaluation als Indikator für die mangelnde Akzeptanz des konkreten Evaluationsvorhabens und der Art und Weise seiner Umsetzung angesehen werden. Wie Strittmatter (2001) sicher nicht nur für die Situation in der Schweiz ausführt, ist mittlerweile in der Lehrerschaft mit einer gewissen «Übersättigung» durch eine Vielzahl von Reformversuchen und den dabei gemachten Erfahrungen und Frustrationen zu rechnen. Häufig ist für die freiwillige Teilnahme an einer Untersuchung ein gewisser «Leidens-Lösungs-Druck» (Strittmatter, 2001) erforderlich. Möglicherweise war auch für die Bereitschaft zur Teilnahme an der WALZER-Befragung die Erwartung ausschlaggebend, in noch weitergehendem Masse von der Untersuchung zu profitieren. Dafür spricht, dass ein nicht unerheblicher Teil der Befragten Interesse

bekundet hatte, mit ihren Schulen und Klassen an weiterführenden Untersuchungen teilzunehmen.

Verständnisschwierigkeiten. Selbstkritisch ist zu vermerken, dass die erste, unter grossem Zeitdruck erstellte Leistungsrückmeldung die Erwartungen vieler Lehrkräfte nicht erfüllte. Es gab Missverständnisse und Unklarheiten bei der gewählten Darstellung, die zu Recht kritisiert wurden. Aus heutiger Sicht ist insbesondere zu bemängeln, dass für die Leistungsrückmeldungen bei MARKUS eine für Lehrkräfte zu wenig aussagekräftige *soziale Bezugsnorm* (Vergleich des Ergebnisses der eigenen Klasse mit denen des gesamten Bildungsgangs) verwendet wurde. Es wäre zielführender gewesen, sich wie bei TIMSS und PISA an *Kompetenzstufen* zu orientieren, die den Lehrkräften genauere Aufschlüsse über vorhandene Leistungsdefizite ihrer Klassen vermittelt hätten. Die dafür erforderlichen Analysen des Mathematiktests (Stichwort: Rasch-Modell) führten jedoch nicht zu dem gewünschten Ergebnis, d.h. es liessen sich keine befriedigenden Kompetenzstufen entwickeln.[3]

Von einer repräsentativen Erhebung «der» Schulleitungen kann somit nicht die Rede sein. Vielmehr dürfte die Stichprobe der befragten Schulleiter/innen durch einen Bias charakterisiert sein: durch die Bereitschaft, sich trotz knapper Zeit mit einem Fragebogen zu beschäftigen, und dies zu einem Thema, das in den Schulen mit sehr viel Aufwand und Organisation verbunden war und inzwischen «abgehakt» schien. Mit einiger Wahrscheinlichkeit handelt es sich – obwohl wir dies nicht empirisch belegen können – um besonders aktive, engagierte, gewissenhafte und pflichtbewusste Schulleiter/innen und/oder solche, die sich für ihre Schulen einen besonderen Nutzen von der Teilnahme an der Untersuchung versprochen haben. Alle folgenden Analysen müssen auf dieser Folie betrachtet werden, und die Ergebnisse sollten nicht auf die Population «der» Schulleitungen verallgemeinert werden.

Trotz der ungesicherten Repräsentativität vermitteln die Ergebnisse der Befragung ein interessantes und recht facettenreiches Bild darüber, wie ein für die Schulen bedeutsames, vom zuständigen Ministerium mit grossem Nachdruck betriebenes Evaluationsvorhaben in den Schulen letztendlich ankommt. Auch wenn die Resonanz der Studie entgegen der im Vorfeld zu beobachtenden Stimmungslage aufgrund der vorliegenden Befragungsergebnisse als nicht übermässig hoch zu veranschlagen ist und die Ergebnisse insgesamt gesehen eher erwartungsgemäss und zufriedenstellend ausfielen, sieht doch ein beträchtlicher Prozentsatz der Schulen, deren Schulleiter/innen an der Befragung teilgenommen haben, Anlass für Veränderungen. Dies führte dazu, dass an diesen Schulen, die allerdings nur etwa 10% aller in Frage kommenden rheinland-pfälzischen Schulen ausmachen, eine Reihe von Massnahmen und Aktivitäten des schulischen Qualitätsmanagements eingeleitet oder zumindest ausgebaut wurden.

Zur Vorhersage des Umfangs, in dem diese Aktivitäten und Massnahmen eingesetzt werden, wurde ein einfaches Bedingungsmodell zugrunde gelegt, das zumindest in einigen Teilaspekten bestätigt werden konnte. Die Ergebnisse lassen

sich so interpretieren, dass das in der Evaluation erreichte Leistungsergebnis einen Änderungsdruck in den Schulen und bei den Schulleitungen erzeugt, der dann zur Einleitung von Massnahmen und Aktivitäten des schulischen Qualitätsmanagements führt. Überraschenderweise hat dabei allerdings ein Merkmal, dem wir einen besonderen Stellenwert beigemessen haben, das vom Schulleiter wahrgenommene *Innovationspotenzial* der Schule, keine bedeutsame Rolle gespielt.

Wie ist das zu erklärten? Vielleicht war die Erfassung und Messung dieses Merkmals suboptimal (Deckeneffekte, Antwortverzerrungen durch soziale Erwünschtheit), vielleicht schätzen die Schulleiter/innen die Situation an ihren Schule auch nicht korrekt ein. Oder: Die vorhandene Innovationsbereitschaft wird erst dann wirksam, wenn sie gebündelt, kanalisiert oder in sonst irgendeiner Weise nutzbar gemacht wird. Dies verweist darauf, dass bei derartigen Faktoren im Grunde keine einfachen Beziehungen zu erwarten sind, sondern vielfältige Wechselbeziehungen und Interaktionen. Bei der Erfassung der Massnahmen der Qualitätssicherung war danach gefragt worden, ob diese *infolge von MARKUS* ausgebaut oder initiiert wurden. Es ist denkbar, dass die aus MARKUS resultierenden Ergebnisse als zu wenig relevant angesehen werden bzw. unterhalb der Schwelle liegen, die überschritten werden muss, um das vorhandene Innovationspotenzial anzusprechen. Möglicherweise fehlt hier einfach ein wichtiges Glied in der Kette von der Rückmeldung des Ergebnisses zur Entstehen einer handlungswirksamen Veränderungsbereitschaft. Dies entspräche einem multiplikativen Modell verschiedener Bedingungsfaktoren.

Ziel weiterer Analysen muss es sein, das Geflecht von Faktoren und ihr Zusammenwirken genauer aufzuhellen. Dies könnte zum einen durch den Einsatz komplexerer statistischer Modelle der Datenanalyse (Strukturgleichungs- oder Kausalmodelle) geschehen. Im vorliegenden Fall sind solche Methoden mit Ausnahme exploratorischer Pfadanalysen (Noonan & Wold, 1988) aufgrund der geringen Stichprobengrösse nicht anwendbar. Sinnvoll und aussichtsreich dürfte es zum anderen aber auch sein, bei künftigen Studien mit Hilfe von qualitativen Methoden (Interviews, Fallstudien) Vorgänge der Rezeption und Nutzung von Rückmeldungen im Detail zu rekonstruieren, um so das Potenzial von Evaluationsstudien zukünftig stärker auszuschöpfen als bisher.

Anmerkungen

1 Wirkungsanalyse der Leistungsevaluation: Zielerreichung, Ertrag für die Bildungsqualität der Schule und die Rückmeldung von Evaluationsergebnissen.
Das Projekt wurde von der Deutschen Forschungsgemeinschaft im Rahmen des DFG-Schwerpunktprogramms «Bildungsqualität von Schule» (Geschäftszeichen HE 1873/2-1) gefördert.
Wir bedanken uns bei H. Ditton und R. Peek, von deren Erfahrungen und Unterstützung das Projekt WALZER erheblich profitiert hat.

2 Die Ergebnisse der Lehrerbefragung werden an anderer Stelle berichtet (Schrader & Helmke, in Druck).

3 Innerhalb des MARKUS-Teams war die Forschungsgruppe Jäger (ZepF Landau) für die Mathematiktests verantwortlich (Balzer & Jäger, 2001), unsere Forschungsgruppe (Helmke, Hosenfeld, Ridder, Schrader) für die Befragungen (Schüler-, Lehrer-, Schulleiterbefragung).

Literatur

Ackeren, I. v. (2002). Von FIMS und FISS bis TIMSS und PISA. Schulleistungen in Deutschland im historischen und internationalen Vergleich. *Die Deutsche Schule, 2,* 157-175.

Ackeren, I. v. & Klemm, K. (2000). TIMSS, PISA, LAU, MARKUS und so weiter – Ein aktueller Überblick über Typen und Varianten von Schulleistungsstudien. *Pädagogik, 12,* 10-15.

Balzer, L. & Jäger, R. S. (2001). Fachleistung Mathematik in MARKUS. *Empirische Pädagogik, 15,* (4), 535-552.

Baumert, J., Artelt, C., Carstensen, C. H., Sibberns, H. & Stanat, P. (2002). Untersuchungsgegenstand, Fragestellungen und technische Grundlagen der Studie. In Deutsches PISA – Konsortium (Hrsg.), *PISA 2000 – Die Länder der Bundesrepublik Deutschland im Vergleich* (S. 11-38). Opladen: Leske + Budrich.

Beck, B. & Klieme, E. (in Druck). DESI – Eine large-scale-Studie zur Untersuchung des Sprachunterrichts in deutschen Schulen. *Zeitschrift für empirische Pädagogik*.

Ditton, H., Arnoldt, B., Babic, B., Bornemann, E. & Zehme, M. (2002a). *Qualitätssicherung in Schule und Unterricht durch Feedbackverfahren*. Beitrag präsentiert am Kongress der Deutschen Gesellschaft für Erziehungswissenschaft, München.

Ditton, H., Arnoldt, B. & Bornemann, E. (2002b). Entwicklung und Implementation eines extern unterstützenden Systems der Qualitätssicherung an Schulen – Quassu. In M. Prenzel & J. Doll (Hrsg.), *Bildungsqualität von Schule: Schulische und ausserschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen. Zeitschrift für Pädagogik. 45. Beiheft* (S. *374-389). Weinheim: Beltz.*

Ditton, H. & Merz, D. (2000). *Qualität von Schule und Unterricht. Bericht über die Voruntersuchung in Bayern.* Eichstätt: Universität Eichstätt.

Ditton, H., Merz, D. & Edelhäusser, T. (2002c). Einstellungen von Lehrkräften und Schulleiter/innen zu zentralen Testuntersuchungen an Schulen. *Empirische Pädagogik, 16,* (1), 17-33.

Helmke, A. (2000). TIMSS und die Folgen: Der weite Weg von der externen Leistungsevaluation zur Verbesserung des Lehrens und Lernens. In U. P. Trier (Hrsg.), *Bildungswirksamkeit zwischen Forschung und Politik* (S. 135-164). Zürich: Rüegger.

Helmke, A. (2003). *Unterrichtsqualität: Erfassen, Bewerten, Verbessern*. Velber: Kallmeyersche Verlagsbuchhandlung.

Helmke, A., Goebel, K., Hosenfeld, I., Schrader, F.-W., Vo, T. & Wagner, W. (in Druck). Zur Rolle des Unterrichts im Projekt DESI. *Empirische Pädagogik*.

Helmke, A., Hosenfeld, I., Schrader, F.-W. & Wagner, W. (2002). Unterricht aus der Sicht der Beteiligten. In A. Helmke & R. S. Jäger (Hrsg.), *Die Studie MARKUS – Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext. Grundlagen und Perspektiven* (S. 325-411). Landau: Verlag Empirische Pädagogik.

Helmke, A. & Jäger, R. S. (Hrsg.). (2002). *Die Studie MARKUS – Mathematik-Gesamterhebung Rheinland-Pfalz: Kompetenzen, Unterrichtsmerkmale, Schulkontext*. Landau: Verlag Empirische Pädagogik.

Helmke, A. & Schrader, F.-W. (2001). Von der Leistungsevaluation zur Unterrichtsentwicklung. In R. Silbereisen & M. Reitzle (Hrsg.), *Psychologie 2000. Bericht über den 42. Kongress der Deutschen Gesellschaft für Psychologie in Jena* (S. 594-606). Lengerich: Pabst.

Institut für Schulentwicklungsforschung. (1999). *IFS – Schulbarometer. Ein mehrperspektivisches Instrument zur Erfassung der Schulwirklichkeit.* Dortmund: IFS-Verlag.

Klemm, K. & Schratz, M. (2002). Leistungstests und Schulentwicklung. *Journal für Schulentwicklung, 2,* 4-8.

Klieme, E., Baumert, J. & Schwippert, K. (2000). Schulbezogene Evaluation und Schulleistungsvergleiche. Eine Studie im Anschluss an TIMSS. In H. G. Rolff, W. Bos, K. Klemm, H. Pfeiffer & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung* (Vol. 11, S. 387-420). Weinheim: Juventa.

Kohler, B. (2002). Zur Rezeption von TIMSS durch Lehrerinnen und Lehrer. *Unterrichtswissenschaft, 30,* (2), 158-188.

Lehmann, R. H., Gänsfuss, R. & Peek, R. (2000a). *Qualitätsuntersuchung an Schulen zum Unterricht in Mathematik. Zwischenbericht.* Berlin: Humboldt Universität.

Lehmann, R. H., Husfeldt, V. & Peek, R. (2001). Lernstände und Lernentwicklungen im Fach Mathematik – Ergebnisse der Hamburger Untersuchung (LAU) in den Jahrgangsstufen 5 und 6. In G. Kaiser, N. Knoche, D. Lind & W. Zillmer (Hrsg.), *Leistungsvergleiche im Mathematikunterricht* (S. 29-50). Hildesheim: Franzbecker.

Lehmann, R. H., Peek, R. & Gänsfuss, R. (1997). *Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Klassen an Hamburger Schulen. Bericht über die Untersuchung im September 1996.* Hamburg: Behörde für Schule, Jugend und Berufsausbildung, Amt für Schule.

Lehmann, R. H., Peek, R., Gänsfuss, R., Lutkat, S., Mücke, S. & Barth, I. (2000b). *Qualitätsuntersuchungen an Schulen zum Unterricht in Mathematik (QuaSUM).* Potsdam: Ministerium für Bildung, Jugend und Sport des Landes Brandenburg (MBJS).

Moser, U. & Rhyn, H. (2000). *Lernerfolg in der Primarschule. Eine Evaluation der Leistungen am Ende der Primarschule.* Aarau: Sauerländer.

Noonan, R. & Wold, H. (1988). Partial least squares path analysis. *Educational research, methodology, and measurement. An International Handbook,* 710-716. Elmsford, NY: Pergamon Press.

Peek, R. (1999). Schulmonitoring und Schulentwicklung. Eine Untersuchung zum Beitrag externer Evaluation für die Entwicklung von Schulen. Unveröffentlichtes Manuskript.

Peek, R. (2000). *Erfahrungen von Brandenburger Lehrerinnen und Lehrern mit QuaSUM. Fragebogenerhebung.* Berlin: Humboldt Universität.

Peek, R. (2001). Die Bedeutung vergleichender Schulleistungsmessungen für die Qualitätskontrolle und Qualitätsentwicklung von Schulen und Schulsystemen. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 323-336). Weinheim: Beltz.

Rolff, H.-G. (2001). Was bringt die vergleichende Messung von Schulleistungen für die pädagogische Arbeit in Schulen? In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 337-352). Weinheim: Beltz.

Schrader, F.-W. & Helmke, A. (in Druck). Von der Evaluation zur Innovation? Die Rezeptionsstudie WALZER: Ergebnisse der Lehrerbefragung. *Empirische Pädagogik.*

Stamm, M. (2002). Evaluation und ihre Folgen: Eine unterschätzte pädagogische Herausforderung. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 2,* 181-196.

Strittmatter, A. (2001). Bedingungen für die nachhaltige Aufnahme von Neuerungen an Schulen. *Journal für Schulentwicklung*, 5, 58-66.

Anhang Abb 1

1.6 Wie ist das Diagramm mit den Leistungsergebnissen Ihrer Klasse zu lesen?
Anhand des nachfolgenden Beispieldiagramms werden die wichtigsten Punkte kurz erklärt. Weiter gehende Hinweise zur Interpretation werden für drei weitere Beispielklassen im Anhang gegeben.

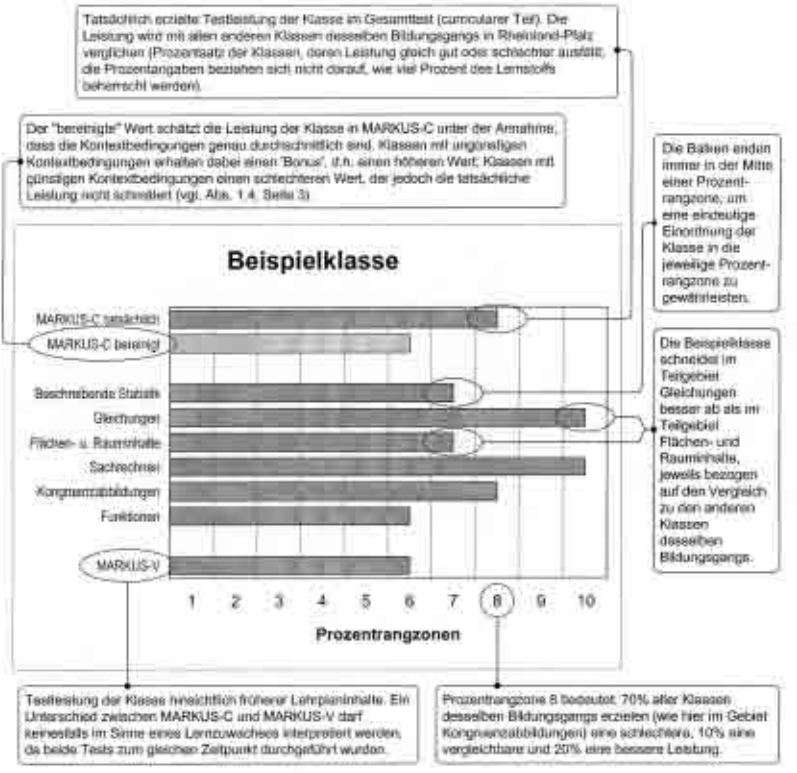

Abb. 1: Erläuterung der Darstellung der Leistungsergebnisse

Bitte beachten Sie bei der Interpretation der Ergebnisse jeweils, in welcher Reihenfolge und in welchem Umfang die Stoffgebiete unterrichtet wurden. Weiter muss auch der Kenntnisstand der Schüler zu Beginn des Schuljahres mit einbezogen werden. Dieser kann aus der MARKUS-Erhebung im Mai 2000 nicht abgeleitet werden. Vielmehr müssen die Lehrkräfte als die Experten für ihre eigenen Klassen hier auf eigene Informationen zurückgreifen.

Anhang Abb 2

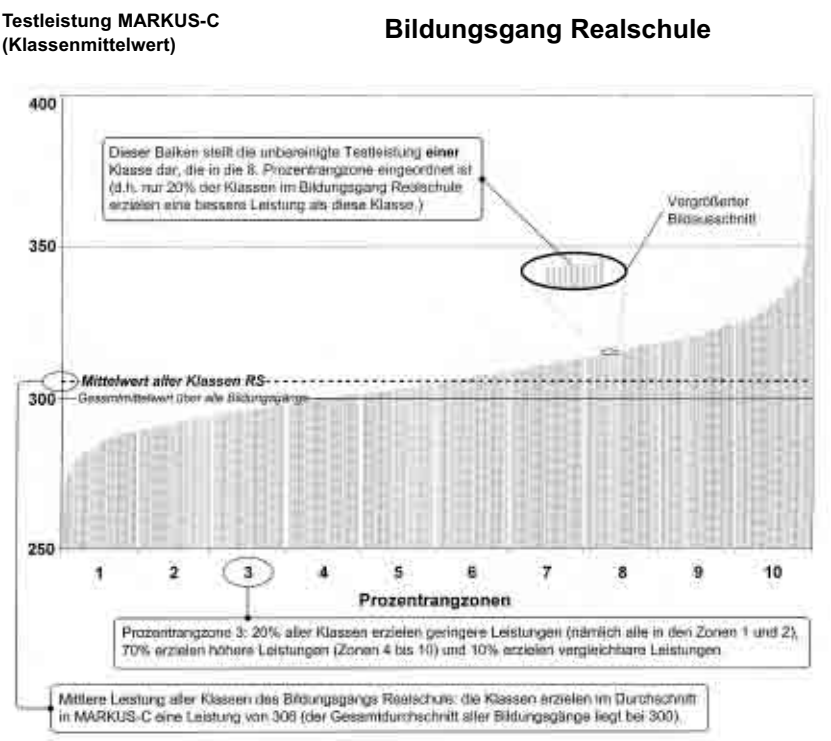

Abb. 3: Testleistung: Zusammenhang zwischen Prozentrangzonen und Rasch-Metrik auf Klassenebene

1.9 Wie sind die Leistungen der rheinland-pfälzischen Schülerinnen und Schüler bei MARKUS-T verglichen mit der TIMSS-Stichprobe einzuordnen?
Durch das Einbeziehen einer Grundmenge von Aufgaben aus TIMSS in MARKUS kann der Frage nachgegangen werden, auf welchem Niveau die Leistungen rheinland-pfälzischer Schülerinnen und Schüler verglichen mit der ursprünglichen Untersuchung von TIMSS einzuordnen sind.

Auch wenn ein solcher Vergleich aus verschiedenen Gründen nur eingeschränkt möglich ist, deuten die Daten darauf hin, dass die in MARKUS untersuchten Schüler des Landes Rheinland-Pfalz (RP) insgesamt besser abgeschnitten hätten als seinerzeit die Schüler der ursprünglichen deutschen TIMSS-Stichprobe. Auf der Ebene einzelner Bildungsgänge sind die Leistungen bei MARKUS sowohl im Bildungsgang Gymnasium als auch im Bildungsgang Realschule in RP bedeutsam besser. Die Leistungen in Hauptschule G bei MARKUS entsprechen denen der Gruppe «Hauptschule» bei TIMSS, die Klassen des

Bildungsgangs Hauptschule A schneiden besser ab als erstere und erreichen die durchschnittliche Leistung der deutschen Gesamtstichprobe von TIMSS. Umfangreichere Informationen zu diesem Vergleich finden Sie im Ersten Ergebnisbericht MARKUS unter http://www.rhrk.uni-kl.de/~zentrum/markus/markus_materialien.html.

2 Kontext

2.1 Welche Bedeutung hat der Kontext ?

Unterricht und Unterrichtserfolg werden in vielfältiger Weise durch Kontextmerkmale der Schüler (z.B. die häusliche Lernumwelt) und der Klassen (z.B. Jungenanteil) beeinflusst. Will man eine faire Beurteilung der erreichten schulischen Leistungen vornehmen, müssen die Kontextbedingungen des Lernens angemessen berücksichtigt werden. Grundsätzlich gilt, dass die Bedeutung einzelner Merkmale nicht überschätzt werden darf. Erst aus der Summe der Merkmale ergibt sich eine angemessene Einschätzung des Kontextes.

Die Kontextmerkmale beeinflussen das Lernen und die Leistung überwiegend nicht direkt, sondern kennzeichnen ein Gefüge (familiärer Hintergrund, Schüler-, Klassen- sowie Schulbesonderheiten), das sich in der Erhebung MARKUS als lernerfolgsbegünstigend oder -erschwerend erwiesen hat.

Anhang Abb 3

In dieser Rückmeldung werden nur diejenigen Merkmale als Kontext bezeichnet, die Lehrkräfte im Unterricht nicht beeinflussen können. Weitere für den Lernerfolg relevante Merkmale wie z.B. Motivation und Lernfreude sind dagegen gleichermaßen Voraussetzung wie Ertrag des Unterrichts. Sie werden deshalb nicht zum Kontext gerechnet und im Unterrichtsprofil (persönliche Rückmeldung an die Lehrkräfte) dargestellt.

Die in MARKUS erfassten Merkmale erlauben eine fundierte Abschätzung des Klassenkontextes, auch wenn die Liste der Merkmale nicht erschöpfend ist. Andere hier nicht erfasste Merkmale (wie z.B. Intelligenzniveau) hängen aber in hohem Masse mit dem erfassten Kontext zusammen.

2.2 Wie wird der Kontext dargestellt?

Der Kontext wurde bereits bei der Berechnung der bereinigten Testleistung berücksichtigt (siehe Abschnitt 1.4). Hier wird für die Klasse ein Profil der Kontextbedingungen dargestellt (siehe Abbildung 5). Dieses Profil ist ebenfalls auf einen Vergleich innerhalb des jeweiligen Bildungsgangs ausgerichtet und gibt die Kontextausprägungen deshalb wiederum in Form von Prozentrangzonen wieder.

Die für diesen Bildungsgang besonders leistungsrelevanten Kontextmerkmale sind durch dunkle Balken hervorgehoben. Da für andere Merkmale als die Ma-

thematik-Testleistung (z.B. Lerninteresse oder Selbstvertrauen der Schüler) und weitergehende Fragestellungen durchaus andere Kontextmerkmale bedeutsam sein können, sind im Profil jedoch stets alle Merkmale aufgelistet.

2.3 Wie ist das Kontextdiagramm zu lesen?

Anhand des nachfolgenden Beispielprofils werden zentrale Begriffe und Konzepte kurz erklärt. Zugleich werden Quellen von Irritationen angesprochen.

Abb. 4: Erläuterung des Kontextdiagramms

L'évaluation et ensuite ? Les résultats d'une enquête portant sur l'accueil de l'enquête WALZER auprès des maîtres principaux.

Résumé

Les résultats obtenus à un test de compétence passé de manière étendue à l'ensemble d'une région (projet MARKUS; Rheinland-Pfalz, Allemagne) engendrent des effets de feedback au niveau des écoles que cet article explore. Dans une enquête auprès des maîtres principaux des collèges (N=52) différents aspects du feedback que l'école a reçu, de même que les attitudes au sujet de l'évaluation et l'innovation, ainsi que les activités de maintien de la qualité scolaire ont été explorés. Pour la plupart des écoles, le feedback des résultats est considéré comme une information importante, suggérant un besoin de changement et conduisant à différentes activités de gestion de la qualité. Des analyses ultérieures montrent que des activités de gestion de la qualité peuvent être prédites par différents jugements et attitudes des maîtres principaux.

Valutazione e dopo? Risultati di un'inchiesta svolta presso i direttori d'istituto nell'ambito del progetto WALZER

Riassunto

L'inchiesta presso i direttori d'istituti scolastici mirava a verificare la ricezione di informazioni sulle prestazioni degli allievi derivate da uno studio a livello regionale (MARKUS-Studie nella regione della Renania Palatinato) e nel contempo a valutare le trasformazioni da esse indotte. Ai direttori (N = 52) sono state sottoposte domande su diversi aspetti di queste informazioni, relative ad una valutazione generale e pure concernenti diverse attività tese ad assicurare la qualità. I risultati mostrano che le informazioni in questione vengono recepite dalla maggioranza delle scuole come stimoli per promuovere i necessari cambiamenti e che in effetti portano alla messa in atto di diverse misure di gestione della qualità. Ulteriori analisi evidenziano che l'applicazione di tali misure è prevedibile sulla base delle valutazioni e delle opinioni dei direttori d'istituto.

Evaluation – and than what? The results of a survey of school principals in the context of the WALZER reception study.

Summary

The article deals with the effects of feedback about achievement test results school classes have obtained in a state-wide evaluation (project MARKUS in Rheinland-Pfalz, Germany). In a survey of school principals (N = 52), different aspects of feedback schools have received, attitudes towards evaluation and in-

novation, and activities of quality management in schools are explored. Feedback about achievement test results is seen as an important information by most schools suggesting a need for change and leading to various activities of quality management in schools. Further analyses show that quality management activities can be predicted by different judgements and attitudes of school principals.