Evaluation im Spiegel ihrer Nutzung: Grande idée oder grande illusion des 21. Jahrhunderts?[1]

Margrit Stamm
Institut für Bildungs- und Forschungsfragen IBF

1. Einleitung: Das Problem

Die vorliegende Arbeit befasst sich mit der Frage, ob und inwiefern die Gesellschaft von Evaluation lernen kann und lernen will. Auf den ersten Blick erscheint eine solche Frage allerdings als überflüssig, da sich Evaluation ja geradezu als Wissenschaft für die Praxis versteht und deshalb die Bereitstellung nützlichen Wissens für Abnehmer- und Anwendersysteme als selbstverständliche Leistung erachtet. Dieses Selbstverständnis findet denn auch in der gesellschaftlichen Akzeptanz seinen Niederschlag, ist Evaluation doch unbehelligt zu einer Standardaufgabe, zu einer *grande idée* nationaler, regionaler und privater Bildungssysteme emporstilisiert worden. Dabei überwiegt der allgemeine Konsens, Evaluation habe Anstöße zur Entwicklung, Optimierung und Veränderung pädagogischer Praxis zu übernehmen. Ein solcher Konsens basiert auf der unhinterfragten Annahme, ein forcierter Einsatz von Evaluation führe automatisch zu einer Qualitätssteigerung von Bildungsleistungen und Bildungsangeboten. Entsprechend hat sich Evaluation bis heute kaum darüber ausweisen müssen, ob sie ihre intendierten Wirkungen tatsächlich erzielt, ob das, was sie produziert, auch für die Adressaten gut ist und von ihnen genutzt wird und sie ihre Leistungsfähigkeit anhand solchermaßen bereitgestellter Verwendbarkeit unter Beweis stellt oder ob sie möglicherweise lediglich eine Bevormundungspraxis darstellt, die sich nur wegen ihrer Wirkungslosigkeit hat institutionalisieren können. Auch wenn die Eingangsfrage – ob und inwiefern die Gesellschaft von Evaluation lernen kann und lernen will – lediglich mit einer beschwichtigenden Antwort und einem Verweis auf gelungene Beispiele quittiert wird, dann drängen sich trotzdem weitere Fragen auf: Warum ist man so sicher, dass Evaluation tatsächlich und in jedem Fall zur Qualitätssteigerung von Bildungsleistungen beiträgt? Warum glaubt man so leichtfüßig an Evaluation, wenn man ihre tatsächlichen Wirkungen doch nicht kennt? Und warum fragt sich Evaluationsforschung

1 Dieser Beitrag ist aus der Habilitationsschrift der Verfasserin mit dem Titel „Evaluation und ihre Folgen für die Bildung: eine unterschätzte pädagogische Herausforderung“ hervorgegangen, die im September 2002 an der Universität Fribourg (Schweiz) am Departement für Erziehungswissenschaften eingereicht worden ist.

nicht selbst, ob sie vielleicht mit hohen Irrtumswahrscheinlichkeiten belegt ist und ihr Nützlichkeitspostulat – Evaluation legitimiere sich durch die Nutzung ihrer Befunde – möglicherweise nur ein „unhaltbarer Mythos" (Drerup 1982: 104) darstellt? Evaluation und ihre Nutzung: vielleicht doch eher eine *grande illusion?*

Dieser Aufsatz gibt einen Einblick in die Frage, ob und wie Evaluationswissen genutzt und weiterverarbeitet wird, welche Bedingungen Nutzungsprozesse begünstigen und falsche Nutzung verhindern. Dabei wird zuerst die Rolle der Erziehungswissenschaft in der Evaluationsforschung thematisiert (2) und PISA als Musterbeispiel einer ‚erfolgreichen' Evaluation charakterisiert (3). Anschließend werden diese Ausführungen in den Ergebnissen einer Feldstudie gespiegelt, die im Rahmen einer empirischen Typologie (4) zu vier verschiedenen Nutzungstypen von Evaluation geführt hat: zur ‚Blockade', zum ‚Alibi', zur ‚Reaktion' und zur ‚Innovation' (5). Diese Ergebnisse münden in eine insgesamt zwiespältige Bilanz, die nicht umhin kommt, Evaluation – auf der Folie der Nutzungsfrage – als problematisches Geschäft zu bezeichnen, ihr aber gleichzeitig die Chance eröffnet, zukünftig als knapper eingesetztes Gut mehr und gezieltere Wirksamkeit als Bildungsprozess zu entfalten (6).

2. Die Rolle der Erziehungswissenschaft in der Evaluationsforschung

Die Frage nach der Nutzung von Evaluationswissen trifft die Erziehungswissenschaft relativ unvorbereitet. Und dies, obwohl die Vertrautheit mit Evaluation dazu geführt hat, sie als selbstverständlichen Teil der Pädagogischen Psychologie zu bezeichnen (vgl. Krapp/Weidenmann 2001; Rost 2001) und auf ihre unverzichtbare Rolle in der Steuerung oder Entwicklung unserer Bildungssysteme zu verweisen (Altrichter/Posch 1997; Burkard/Eikenbusch 2000; Helmke 2000; Lange 1999; Tillmann/Vollstädt 2001). Allerdings mangelt es auch nicht an Abhandlungen, welche die praktische Evaluationsrelevanz herausstreichen und davor warnen, Evaluation zum Selbstzweck werden zu lassen (Beywl 1988, 1999; Guba/Lincoln 1981; Hellstern/Wollmann 1983; Kordes 1983; Wottawa/Thierau 1998). Merkwürdig bleibt jedoch, dass die Folgen von Evaluation nicht häufiger thematisiert werden, obwohl gerade der erziehungswissenschaftliche Bereich Evaluation mit besonderen Erwartungen an die Umsetzung verbindet. Folgedessen müsste vermehrt Unbehagen über die Wirksamkeitsfrage formuliert oder die Forderung aufgestellt werden, Evaluation habe ihre Effektbehauptungen nachzuweisen und aufzuzeigen, ob und wie Evaluationsanalyse und nachfolgende Intervention miteinander verbunden werden (Oelkers 1997; Pekrun 2002; Terhart 2002). Versteh- und interpretierbar wird dieses fehlende Unbehagen erst, wenn man bedenkt, dass eine pädagogische Evaluationstradition fehlt, obwohl sich Probleme des Transfers von Wissen in die Praxis stellen, seitdem die Erziehungswissenschaft als Pädagogik versucht, sich als eigene Wissenschaft zu begründen, so wie dies Herbart oder Schleiermacher getan haben. Gerade aus solchen Gründen wäre Erziehungswissenschaft prädestiniert, sich erfolgreich in die Diskussion um eine Verwendungstheorie der Evaluationsforschung einzumischen. Angesichts der zersplitterten, disziplinär organisierten Eva-

luationsbemühungen in den Sozialwissenschaften dürfte dies gar nicht so einfach sein. Zwar werden immer wieder Stimmen laut, die eine verbindende Perspektive zwischen den eher objektivistischen politikwissenschaftlichen und den qualitativ orientierten Evaluationsmodellen des pädagogisch-psychologischen Bereichs fordern, doch scheint gerade die Zunahme der Akzeptanz qualitativer, der Handlungsforschung nahestehender Modelle nur zögerlich vonstatten zu gehen (Gather Thurler/Huberman 1993). Dies liegt nicht zuletzt in ihrer mäßigen Akzeptanz durch die Bildungspolitik, welche ihnen nur geringe Außenwirksamkeit attestiert. Dessen ungeachtet bleibt die Feststellung, dass Evaluation weitgehend ohne theoriegeleitete Metaevaluationen betrieben wird, ähnlich wie man sie aus der Therapieforschung kennt (Grawe/Donati/Bernauer 1993), so dass bis heute ungeklärt ist, welche Methoden und Modelle besondere katalytische, auf Umsetzung bezogene Wirksamkeit entfalten können.

Geht man von der Satzung aus, dass Schulleistungsvergleichsstudien eine neue und vorrangige Aufgabe nationaler Bildungssysteme werden, Evaluation jedoch ein vorwiegend pädagogisches Geschäft sein soll, so müsste die Erziehungswissenschaft in diesem Bereich künftig gehaltvoll mitreden. Dann wird sie allerdings nicht darum herum kommen, sich mit ihrem alten, in neuem Zusammenhang gestellten und bisher ungelösten Problem zu beschäftigen, nämlich ob und wie theoretisches Wissen in die pädagogische Praxis überführt werden kann, respektive unter welchen Bedingungen Gesellschaft und Bildungspolitik bereit sind, von Evaluation zu lernen. In der Konsequenz bedeutet dies, dass die Erziehungswissenschaft, will sie sich erfolgreich an der Weiterentwicklung der Evaluationstheorie beteiligen, mit Vorteil die breiten Forschungserfahrungen aus den achtziger Jahren zur Verwendung sozialwissenschaftlichen Wissens in außerwissenschaftlichen Kontexten berücksichtigt (Beck/Lau 1983; Beck/Bonß 1989a, b; Drerup 1979, 1982, 1989; Drerup/Terhart 1990; Evers/Nowotny 1989; Weber 1994). Die Wissensverwendungsforschung zeigt, dass von der Erkenntnis zur Verwendung wissenschaftlichen Wissens in pädagogischen Praxiszusammenhängen eine große, ernüchternde Kluft besteht und dass es keinen gradlinigen Transfer von der Wissenschaft in die (pädagogische) Praxis gibt. Es empfehlen sich jedoch auch Anleihen bei der angloamerikanischen *Research on evaluation utilization* (Alkin/Coyle 1988; Cousins/Leithwood 1986, 1993; Daillak 1982; Heller 1986; Patton 1986, 1997; Siegel/Tuckel 1985; Weiss 1972c, 1982) und auch bei der wissenschaftlichen Begleitforschung der Bildungsreform der siebziger und achtziger Jahre. Im Unterschied zu heute stand damals zwar nicht die Überprüfung der Qualität von Bildungsleistungen im Mittelpunkt, sondern die Analyse der Bedingungen, Formen und Folgen von Reformmaßnahmen (Nuthmann 1983), doch lassen sich die wesentlichsten Erkenntnisse trotzdem auf die aktuelle Situation übertragen: Im Ergebnis zeigt die wissenschaftliche Begleitforschung, dass sich die Annahme ‚Je besser die methodische Perfektion, desto höher die Nutzung' als falsch, mit umgekehrten Vorzeichen jedoch als richtig erweist: Je höher die methodische Qualität und Raffinesse einer Evaluation und je ausgeprägter die theoretische Steuerung und Erkenntnisgewinnung, desto geringer ist die Chance, dass die Ergebnisse Entscheidungsfindungen oder andere Folgeaktivitäten beeinflussen können. Da methodische Perfektion letztlich die Abnehmerbedürfnisse nur schwer befriedigen und Lernen nicht garantieren kann, gerät sie in Gefahr, zu einer der hauptsächlichen Quellen des Nutzendefizits zu werden.

Ein differenzierter Blick in die Bezugsfelder der *Research on evaluation utilization* verweist jedoch auf eine Schwierigkeit: Es besteht eine große Diskrepanz zwischen den vielfältigen theoretischen Konzepten über Faktoren, welche Evaluationsnutzung beeinflussen (Shulha/Cousins 1997; Weiss 1972a, b, 1998) und den wenigen systematischen Feldstudien zur Nutzung von Evaluationsbefunden (wie etwa die Interviewstudien von Dickey 1981, die viel beachtete Studie von Weiss/Bucuvalas 1980 oder die naturalistischen Fallstudien von Alkin 1980 oder Greene 1987). Diese Studien werfen zwar das Licht auf eine offenbar recht umfangreiche Anzahl nicht erfolgreicher Evaluationen, die Nutzung in der Tat als Illusionsformel entlarven, doch macht die Bilanzierung des Status quo der amerikanischen Forschungsliteratur gleichzeitig klar, dass die Theorie der Evaluationsnutzung bei weitem die Empirie übersteigt und deshalb keine genügend große empirische Basis für solche Urteile vorliegt. Letztlich konkretisiert sich das Problem der Evaluationsnutzung damit in seinen unzulänglichen Erklärungsversuchen, so dass nichts weiter bleibt, als sich „unterhalb solcher Verlegenheiten zu konsolidieren" (Luhmann/Schorr 1982: 7) und bei der konkreten Evaluationspraxis anzusetzen. Dies geschieht anhand der Schulleistungsvergleichsevaluation PISA und der Feldstudie über Schweizer Bildungsevaluationen.

3. PISA: Musterbeispiel einer erfolgreichen Evaluation?

Die Schulleistungsvergleichsevaluation PISA ist es, welche die Eingangsfrage von ihrer vordergründigen Überflüssigkeit entbindet: Sie zeichnet beispielhaft die Problematik dieses Aufsatzes nach und markiert den Idealtypus einer Evaluation, in welchem die Problematik der Rezeption, des Transfers und der Nutzung von Evaluationswissen aufscheint – gewissermaßen als Musterlektion von Evaluationsnutzung. Misst man den Erfolg von PISA nämlich ausschließlich am Ausmaß ihrer Rezeption, so kann bereits ein gutes Jahr nach der Veröffentlichung der Ergebnisse konstatiert werden, dass es sich um die Evaluation mit dem bisher größten *Impact* überhaupt handelt und sie deshalb als ‚erfolgreichste' Evaluation aller Zeiten zu bezeichnen wäre. Die positive Rezeption gilt grundsätzlich auch für den methodologischen Bereich und ihre bildungstheoretische Einbettung, obwohl durchaus auch legitimationskritische Fragen und Kritik laut werden (Benner 2002; Hagemeister 1999, 2000a, b; von der Groeben 2000, 2002). Bei näherem Zusehen hin erweisen sich die Folgen von PISA allerdings als problematisch. Denn die Evaluationsbefunde werden selten in dem Kontext übernommen, in welchem sie produziert worden sind, vielmehr werden sie häufig transformiert, reduziert und so umgebogen, dass sie bildungspolitische Standpunkte legitimieren können. PISA wird damit zum Abbild der großen Verflochtenheit von Erkenntnis und Interesse, die sich in der Indienstnahme durch politisch-praktische Interessen, spezifische Verwendungszusammenhänge und Verwendungszwecke äussert. Die ‚Folgen von PISA' zeigen damit exemplarisch auf, dass sich Evaluationen keinesfalls lediglich durch das Ausmaß der Nutzung ihrer Ergebnisse für nachfolgende Entscheidungen, durch ausgereifte Methodologie oder technische Perfektion legitimieren lassen oder man ihr Ausbleiben als alleinige Quelle eines möglichen Nutzungsdefizits interpretieren

kann. Nutzungsprobleme artikulieren sich als Problemlagen in zwei pädagogisch bisher kaum gewürdigten Bereichen: Erstens als durch die Bildungspolitik generierte Transformationsproblematik, welche die Ergebnisse nicht so verwendet, wie sie von der Forschung erzeugt worden sind; zweitens als Wissenserschließungsproblematik, die sich in der fehlenden Interpretationskompetenz der Daten durch die Abnehmersysteme manifestiert und einen im Sinne der Evaluatorinnen und Evaluatoren liegenden Umgang mit den Daten erschwert. Selektive Verwendung einerseits und fehlende Interpretationskompetenz andererseits verweisen jedoch auf einen dritten Problembereich, auf die Frage nämlich, *wer* denn die Vermittlungsleistungen zu erbringen hätte und wie diese zu gestalten wären. Für Weinert (2002) ist die Antwort klar: Die didaktisch-bildungsstrategische Aufbereitung von evaluativ gewonnenem Wissen und auch die Verbreitung und Kommunikation der Befunde gehört in den Verantwortungsbereich der Evaluationsforschenden selbst und bildet einen wesentlichen Teil ihrer Berufsaufgabe. Wenn demzufolge Evaluatorinnen und Evaluatoren für Anwendungsfragen zukünftig mehr Verantwortung übernehmen müssen, dann können sie sich nicht mehr ausschließlich als wissengenerierende Fachleute für eine Gelehrtenprofession verstehen:

> „Entscheidend für die Nützlichkeit des Wissens ist nicht nur die wissenschaftliche Solidität und Ergiebigkeit, sondern auch und vor allem die Kommunikation der ermittelten Befunde an Ministerien, Schulen, Lehrer und wenn nötig und möglich auch an Schüler, Eltern und interessierte Öffentlichkeit." (ebd.: 359)

Gerade im Falle von PISA zeigt sich dieses Kommunikationsproblem verschiedentlich, zu dem die Forschenden selbst ihren Teil dann beitragen, wenn sie Aufklärungsarbeit weitgehend unterlassen oder ihre spärlichen Stellungnahmen einzig an die Forschergemeinde adressieren und persönliche Folgerungen und Erklärungen[2] dessen, wozu Leistungsvergleichsstudien in der Lage sind und wozu nicht, zu wenig umsetzungsverträglich und anwenderorientiert aufbereiten (Terhart 2002: 104ff.). Damit tragen sie maßgeblich dazu bei, dass eine öffentliche Diskussion ebenso fehlt wie die Herstellung von Anwendbarkeit, welche die komplexen Daten mittels Zusatz- und Orientierungswissen transparent macht und eine Kommunikation erlauben würde, die nicht in formelhafte Kausalvermutungen mündet. Möglicherweise sind das Gründe, die für die Oberflächlichkeit und Selektivität der Rezeption mitverantwortlich sind.

Belege dafür, dass das Kommunikationsproblem allerdings nicht so eindeutig in den Verantwortungsbereich der Evaluationsforschenden delegiert werden kann, finden sich in den Erkenntnissen der Wissensverwendungsforschung beispielsweise dort, wo die Kluft zwischen Erkenntnis und Gestaltung umschrieben wird (Drerup/Terhart 1990). Sie trägt dazu bei, dass Evaluationsergebnisse immer ungeschützt sind und die Forschenden selbst kaum verhindern können, dass frei zugängliche Evaluationsinformation zu verschiedenen Zwecken missbraucht wird. Diese

2 Die Zurückhaltung ist bislang in der Schweiz größer als in Deutschland, wo Jürgen Baumert kontinuierlich versucht, den Kurs zu korrigieren und Aufklärungsarbeit zu leisten (vgl. beispielsweise: Wieso, weshalb, warum? Ein Gespräch mit Jürgen Baumert, verantwortlich für die TIMS-Studie, und Hermann Lange, PISA-Beauftragter der Kultusministerkonferenz. In: Die Zeit, Nr. 50: 78).

Kluft wird durch die Typik der Evaluation als Auftragsforschung mit spezifischem, teilweise politisch konnotiertem Verwendungsinteresse der Anwendersysteme zusätzlich vergrößert. PISA zeigt es: Die Bildungspolitik behandelt die Ergebnisse stark selektiv, generiert selbst Umsetzungswissen und nimmt Ableitungen und Interpretationen vor, die über das von den Forschenden selbst Intendierte und in der Untersuchungsanlage Inhärente hinausgehen.

Bezogen auf die Problematik der Folgen von Evaluationsuntersuchungen fällt die Bilanz von PISA deshalb eher ernüchternd aus. PISA erzeugt zwar hoch differenzierte Information und weist Schwächen und Stärken der einzelnen Länder in den verschiedenen Kompetenzbereichen spezifisch aus, erzeugt allerdings Überdosen an Rückmeldungen, die ihrerseits wiederum zu wenig Hinweise dazu liefern, wie angesichts der beschränkten Aussagekraft der Ergebnisse mit diesem Wissen umzugehen, respektive wie es zu interpretieren wäre. Solche Voraussetzungen laden geradezu zu Verzerrungen oder Reduktionen ein, je nachdem, ob die Ergebnisse eher als Zumutungen, als Konfrontationen oder als Streicheleinheiten verstanden werden. Deshalb sind solche Folgen schlecht genutzte Erkenntnis, die vertiefte Fragen nach dem Nutzen von Evaluation nach sich ziehen, auch wenn sie möglicherweise dazu führen könnten, dass Evaluation in eine Legitimationskrise gerät.

4. Feldstudie und empirische Typenbildung

Die Feldstudie fokussierte auf die Nutzungsprozesse von Evaluation. Als Vollerhebung konzipiert umfasste sie alle in der deutschen Schweiz verfügbaren, von kantonalen Bildungsdepartementen an externe Evaluatorinnen und Evaluatoren in Auftrag gegebenen und zwischen 1995 und 1998 abgeschlossenen Evaluationen. Damit konnte die Varianz des Untersuchungsfeldes sichergestellt und die Vermeidung von Verzerrungen garantiert werden. Um die Fragestellung auf kontrollierte Weise angehen zu können, wurde die Methodologie der vergleichenden Fallanalyse (Kelle/Kluge 1999; Kluge 2000) mit empirischer Typologie als komparativem Ansatz gewählt. Er ist aufgrund seiner vergleichsdimensionenorientierten Basis in der Lage, möglichst alle mit der Nutzung in Zusammenhang stehenden Erscheinungen in ‚dichter Weise' zu beschreiben (Geertz 1999). Über fokussierte Interviews wurden anhand mehrperspektivischer Nutzenschilderungen insgesamt 56 Datensätze erhoben. Parallel dazu wurden Fragebögen eingesetzt, welche quantifizierbare und mit den Interviews kombinierbare Daten als Fallvariablen lieferten (Kuckartz 1999; Witzel 1982). Demzufolge lagen zu jedem Fall qualitative Aussagen und quantifiziere Daten vor. Nach der Codierung des Textmaterials erfolgte die Datenauswertung mit dem Textdatenbanksystem winMAX, das die Verwaltung und Auswertung des umfangreichen Datenmaterials ermöglichte und erlaubte, die individuellen Daten für den Einzelfall, aber auch für den Fallvergleich, darzustellen. In einem ersten Schritt erfolgte die Erarbeitung relevanter Vergleichsdimensionen (Zweck/Ziel und Rezeptions-/Verwendungstyp) mit anschließender Gruppierung der Fälle anhand dieser Merkmale. Hierauf wurden die ermittelten Gruppen hinsichtlich der empirischen Regelmäßigkeiten untersucht und kreuztabelliert, was die Zuordnung der 18 Evaluationen zu einzelnen Feldern sichtbar machte (vgl. Tabelle 1).

Tabelle 1: Vier Typen der Beziehung zwischen Rezeption und Ziel/Zweck

Rezeptions- und Verwendungstyp	Zweck/Ziel		
	Kontrolle	Legitimation	Entwicklung/ Optimierung
direkte Rezeption/ Nutzung nachweisbar	Typus 1* (n = 4)		Typus 3 (n = 5)
direkte Rezeption/ Nutzung unklar direkte Rezeption/ Nutzung nicht vorhanden	Typus 4 (n = 6)	Typus 2 (n = 3)	

Für Typus 1 wurde später die Bezeichnung ‚Reaktion' gewählt, für Typus 2 die Bezeichnung ‚Alibi', für Typus 3 ‚Innovation' und für Typus 4 ‚Blockade'.

Parallel dazu wurde eine Clusteranalyse mit insgesamt 40 Items gerechnet, die aus der Vercodung der Interviewtexte in das empirische Stufenmodell eingeflossen waren (Tabelle 2). Sie bildete gewissermaßen die systematische Auswertung dieser Ähnlichkeitsmatrix. Dabei wurde die Methode des Complete-Linkage-Verfahrens gewählt, ein hierarchisches Verfahren nach agglomerativer Vorgehensweise (vgl. Backhaus 1996: 348). Da mit dieser Clusteranalyse ein Vergleich mit den über das empirische Stufenmodell gewonnenen Ergebnissen stattfinden sollte, ging es darum, eine Clusterlösung zu finden, welche die größte Übereinstimmung aufwies. Letztlich waren es deshalb theoretische Erwägungen, der angestrebte Grad an Differenziertheit und die Größe der einzelnen Cluster, welche dem Entscheid für eine Lösung mit vier Clustern den Vorzug gaben.

Die in Tabelle 2 dargestellten Clusterprofile enthalten für jedes Cluster den prozentualen Anteil, welchen das Cluster an der jeweiligen Merkmalsausprägung aufweist. Diese Matrix ist wichtig für die nachfolgende Interpretation.

Insgesamt lagen nun die Ergebnisse der Clusteranalyse und des empirischen Stufenmodells vor. Nach Rückgriff auf die Originaldaten und die Einzelfallanalysen wurde schließlich der *Lösung des Stufenmodells* der Vorzug gegeben. Anhand der Merkmalskombinationen und Sinnzusammenhänge wurde eine Charakterisierung der Typen vorgenommen, die schließlich zu den Bezeichnungen ‚Blockade', ‚Innovation', ‚Alibi' und ‚Reaktion' führte. Die Kunst der Typenbezeichnung lag dabei darin, eine Formulierung für das Charakteristische in den Einzelfallanalysen zu finden, die zugleich ihr gemeinsam Geteiltes und ihr zu anderen Fällen Differentes darstellen, ohne dass es dabei zu Verkürzungen oder Verzerrungen der Komplexität des Untersuchungsmaterials kommen sollte.

Tabelle 2: Clusterspezifische Merkmalsverteilung

Item	Cluster 1* n Fälle = 5	Cluster 2* n Fälle = 3	Cluster 3* n Fälle = 6	Cluster 4* n Fälle = 4
Überprüfung	33	33	29	44
Legitimation	33	67	21	55
Umsetzung	33	0	60	0
bildungspolitische Bedeutung	67	50	56	78
politische Brisanz	67	33	50	67
Zufriedenheit mit Ergebnissen	44	33	79	33
positive Gesamtbilanz	44	17	100	44
Widerstände	33	0	14	78
Bottom-up Struktur	33	0	78	44
Top-down Struktur	67	100	21	56
direkte Verwendung	33	33	100	100
Veränderungsbereitschaft	44	33	86	78
Unterstützung durch Bildungsdepartement	33	33	86	67
Berufsorientierung ‚Umsetzung'	22	67	29	22
Berufsorientierung ‚Wissenschaft'	78	33	71	78
Zensuransätze	44	78	71	78
Umfang Dissemination	33	70	93	78
sprachliche Qualität	44	67	100	100
Termintreue	67	33	100	33
methodische Glaubwürdigkeit	67	50	100	78
finanzielle Ressourcen	11	17	57	57
Commitment Evaluator/in	11	44	71	56
Reputation Evaluator/in	22	100	71	100
Nützlichkeit/Anwendbarkeit	33	67	50	50
Zufriedenheit mit Nutzung	11	33	93	50
Erfahrungen mit Evaluation	67	67	17	67
Engagement Entscheidungsträger	22	50	100	86

Lesehinweis: Alle Angaben in Prozent.
44% der Evaluationen des ersten Typus weisen eine hohe sprachliche Qualität auf. In Cluster 1 beispielsweise ist nur in 33% der Fälle Widerstand aufgetreten und in 11% bestand Zufriedenheit mit der tatsächlichen Nutzung. Ganz anders hingegen Cluster 4: Hier bestand in 78% der Fälle Widerstand und die Zufriedenheit mit der Nutzung umfasste 50%. Cluster 1 entspricht der ‚Blokkade', Cluster 2 dem ‚Alibi', Cluster 3 der ‚Innovation' und Cluster 4 der ‚Reaktion'.

5. Die vier Typen von Evaluationsnutzung

Das kombinierte Modell empirischer Typenbildung hat im Ergebnis zur Formulierung der vier nachfolgend charakterisierten Typen geführt, auf die hin die Einzelfälle konvergieren. Sie bilden damit die Varianz, mit der Evaluationsnutzung insgesamt verstanden werden kann. Damit stehen vorerst vier Schemata zur Verfügung, die sich methodologisch gesehen als Elemente einer ‚Grammatik' der Rezeptions- und Nutzungsfrage von Evaluation begreifen lassen.

5.1 Der Blockade-Typus

Evaluationen, welche zu gleichen Teilen der Überprüfung, der Kontrolle und der Umsetzung dienen sollen (je 33%), jedoch kaum direkte Nutzung von Evaluationswissen aufzeigen (33%), bilden den Typus der Blockade und zwar deswegen, weil bestimmte Gründe unterschiedlicher Natur eine unmittelbare Umsetzung der Befunde verhindern. Damit trifft die Bezeichnung ‚Blockade' ins Schwarze, weil sie die metaphorische Bedeutung einer Sperre annimmt, welche an der Verhinderung weiterer Aktivitäten beteiligt ist. Entsprechend gering ist die Zufriedenheit mit der Nutzung der Ergebnisse (11%). Gleiches gilt für die Einschätzung der Verwendbarkeit der Ergebnisse (33% Positivwerte). Sowohl das Engagement der Evaluatorenschaft wie auch der Entscheidungskader ist mit 11% resp. 22% relativ gering.

Die Blockade als Nutzungstypus lässt sich vorerst über die beiden Vergleichsdimensionen Überprüfung als Hauptzielsetzung sowie unklare oder nicht vorhandene direkte Rezeption definieren. Da die Entscheidungskader der Blockade-Typik mehrheitlich über einen erziehungswissenschaftlich-pädagogischen Hintergrund verfügen, sind sie in der Lage, das evaluativ gewonnene Wissen autonom zu bewerten. Trotzdem muss ihre Problemerfassung beschränkt bleiben, da Interpretationen nur soweit möglich sind, als dies ihre hierarchischen Organisationsstrukturen mit ausgeprägten Delegationsstrukturen zulassen (Dubs 2000). Solche Strukturen bilden ungeeignete Bedingungen für Lernprozesse und Diskurspolitik, aber geeignete Blockaden, die den Sprung zur Weiterverwendung der Ergebnisse verhindern oder verdunkeln können. Besonders deutlich offenbart sich im Typus der Blockade die Notwendigkeit, die strukturellen Dimensionen und latenten Schattierungen der Makroebene zu berücksichtigen. Dies lässt die Evaluation allerdings zum Teilstück eines politischen Spiels werden. Eine Blockade-Evaluation kann deshalb nur kontextuelle Relevanz erzielen, wenn sie die Bedingungen des Makrokontextes berücksichtigt, was die lokale Verwendung jedoch problematisch werden lässt.

Der Blockade-Typus unterscheidet sich von den drei anderen Typen vor allem in der Dimension der Nutzen- und Anwendbarkeitseinschätzung. Die Erfahrung mit Evaluation und ihr fachliches Know-how erlaubt den Verantwortlichen, aus der Vielzahl von Rezeptionsmöglichkeiten für den Umgang mit diesem ‚anderen', herausfordernden Wissen eine geeignete Variante auszuwählen, nach dem Motto ‚Wie stelle ich die Suppe beiseite, damit wir sie nicht zu heiß löffeln müssen?' Als an Wissenschaft gewöhnte Professionelle sind sie in der Lage, in verwissenschaftlichten Formen zu argumentieren, Befunde zu relativieren und sie Evaluationserkenntnissen aus anderen Untersuchungen gegenüberzustellen. Deshalb werden die Befunde zuerst einmal zur Kenntnisnahme an andere Gremien weitergereicht und über viele Umwege dann schließlich doch in Absichtserklärungen – allerdings mit hinausschiebendem Charakter – überführt, denn die aktuellen bildungspolitischen Zielsetzungen sollen nicht gefährdet werden. Da solche Entscheidungskader die Einschätzung von Nützlichkeit und Anwendbarkeit von Evaluationsinformation immer stark auf ihre politische Verwendungstauglichkeit hin befragen, sind Evaluationen dieses Typs vordergründig gar nicht in der Lage, den für die Auftraggeber relevanten Informationsbedarf zu befriedigen. Damit wird die Blockade zum Paradebei-

spiel der Frage, welchen Voraussetzungen denn Evaluationen genügen müssen und genügen können, um einen höheren Einfluss zu erlangen.

5.2 Der Typus der Innovation

In diesem Typus finden sich Evaluationen, welche eindeutig entwicklungs- und umsetzungsorientierten Charakter haben (60%). Sie lassen sich durch geplante, umfassende und anhaltende Umsetzungsmaßnahmen kennzeichnen, welche den Charakter von Innovationen annehmen. Idealtypische Konturen ergeben sich aus der hohen Qualität der Evaluation (methodische Glaubwürdigkeit: 78%, sprachliche Qualität 100%), den organisationalen Merkmalen, welche Autonomie und Selbstverantwortung fördern (bottom-up Struktur: 78%) und letztlich Dissemination (93%) und tatsächliche Nutzung als Folgen eines fruchtbaren Prozesses erscheinen lassen. Entsprechend hoch ist die Zufriedenheit mit den Evaluationsergebnissen (79%) und deren Nutzung (93%). Die Gesamtbilanz ist vollumfänglich positiv (100%).

Die Hauptkennzeichen der Innovation liegen in der Umsetzungsorientierung und der direkten Rezeption und Nutzung. Sie unterscheidet sich von allen anderen Typiken durch die höchste Zufriedenheit mit der tatsächlichen Umsetzung bei gleichzeitig hohem persönlichem Einsatz der Entscheidungsverantwortlichen und der Evaluationsforschenden. Die eher bottom-up strukturierten Bildungsadministrationen zeigen relativ große Veränderungsbereitschaft, die sich auch in nicht-hierarchischen Beziehungsstrukturen zwischen Auftraggebern und Auftragnehmern äußern. Im Ganzen gesehen zeichnet sich der Innovationstypus durch eine hohe Qualität der Evaluation (sprachliche Verständlichkeit, Termintreue, Methodenwahl) und eine hohe Verwendungstauglichkeit der Befunde aus.

Als Gegenpol des Blockade- und des Alibi-Typus beinhaltet die Innovation viele Aspekte einer idealtypischen Konstruktion von Evaluationsnutzung. Im Gegensatz zur Blockade sind Entscheidungsverantwortliche der Innovation allerdings auch geneigt und in der Lage, den Evaluationsgegenstand selbst zum Problem zu machen und sich persönlich in Lernprozesse einzulassen, Veränderung anzustreben und nicht primär als bedrohlich zu empfinden. Wirklichkeitserfassung wird damit in der Innovations-Typik zur Problemerfassung. Den Entscheidungsträgern, die häufig in Personalunion auch Projektleitung und Auftraggeberrolle innehaben, kommt dabei eine Schlüsselrolle zu. Mit ihrem hohen Engagement schaffen sie für alle Beteiligten ein günstiges, offenes Arbeitsklima und sichern die Überführung der Ergebnisse in gestaltende Maßnahmen. Die günstigen politischen Konstellationen, etwa die Unterstützung durch die bildungspolitische Behörde – auch im Falle personeller Wechsel – tragen dazu bei, dass die Evaluation zur persönlichen Angelegenheit der Verantwortlichen wird und Erkenntnis- und Handlungsinteressen konvergieren können. Zur idealtypischen Konturierung tragen jedoch auch starke Evaluatorenpersönlichkeiten bei, die nicht nur über hohe fachliche Qualität und Feldkenntnis verfügen, sondern auch im Disseminationsprozess eine bedeutsame Rolle spielen und bei der Dateninterpretation unterstützend zur Seite stehen.

Allerdings ist die Innovations-Typik auch gefährdet. Ihr Hauptrisiko besteht in der prozessorientierten Arbeitsweise, die große zeitliche und finanzielle Ressourcen

bindet und den evaluativen Charakter zu Gunsten einer allgemeinen Schulentwicklungsperspektive zu verlieren droht. Die Gefährdung betrifft auch Position und Rolle der evaluierenden Personen, die aufgrund des eher formativ geprägten Evaluationscharakters und der Familiarität mit den Entscheidungsverantwortlichen ihre externe neutrale Position zu Gunsten einer Insiderposition aufgeben und so ungewollt in die Rolle einer verbündeten Person schlüpfen. Innovationstypen, die als externe Evaluationen deklariert werden, machen jedoch nur so lange Sinn, als Evaluatorinnen und Evaluatoren tatsächlich den Expertenstatus beibehalten und sich gegenüber den Auftraggebern und Akteuren auch eindeutig als solche darstellen und solange, als die pädagogische Basis tatsächlich ein Laiengremium bleibt und nicht zu einer selbsternannten Expertengruppe wird. Dazu kommt, dass die Typik der Innovation mit ihren prozessorientierten, bottom-up gesteuerten Evaluationen den bildungspolitischen Behörden weit mehr Sorge bereitet als top-down-orientierte Ansätze, die besser kontrollierbar und deshalb weniger beängstigend sind. Demgegenüber sind die rationalen Taktiken der Reaktion oder die politisch gefärbten Taktiken des Alibis und der Blockade kontrollierbarer und deshalb akzeptierter.

5.3 Die Typik des Alibis

Dieser Typus setzt sich aus Evaluationen zusammen, die in erster Linie legitimierenden Charakter (67%) haben und nur zu 33% direkte Evaluationsfolgen nachweisen können. Die Evaluationen werden deshalb als ‚Alibi' bezeichnet, weil sie kein pädagogisches Erkenntnisinteresse, sondern lediglich der legitimatorische Wert, produziert worden zu sein, miteinander verbindet. Und dies, obwohl sich dieser Evaluationstypus durch den höchsten Anteil dienstleistungsorientierter Evaluatorinnen und Evaluatoren auszeichnet (67%). Alibi-Evaluationen, in allen Fällen (100%) in top-down organisierten Bildungsorganisationen durchgeführt, sind Antworten mit reinem Bestätigungs- und Rechtfertigungscharakter auf von bestimmten Interessengruppen ausgehende Forderungen nach Evaluation. Die Gesamtbilanz fällt entsprechend nur zu 17% positiv aus.

Legitimierende Zielsetzung, keine oder unklare Rezeption und Verwendung – auf diese Kurzformel gebracht bildet die Typik des Alibis den Gegenpart der Innovation. Dort Optimierung, Entwicklung und Nachhaltigkeit, hier Wirkungslosigkeit, Frustration, Bestätigung des schon vorher Gewussten. Im Ganzen gesehen unterscheidet sich das Muster des Alibis von den drei anderen Nutzungstypen durch den Legitimationsdruck, der auf der Evaluation lastet, durch die bescheidene Finanzierung sowie durch bereits vorhandene Erfahrung mit Evaluation. Trotz Handlungsdruck werden Alibi-Evaluationen eher mit Widerwillen durchgeführt, bringen jedoch wegen der bestätigenden Ergebnisse den Verantwortlichen maßive Entlastung, nicht zuletzt auch, weil sie Material für eine ‚Absicherung für alle Fälle' liefern und einen Schlusspunkt unter die bisherige Projektarbeit setzen. Alibi-Evaluationen bringen somit auch Entwarnung. Zwar mit einer ähnlichen Zielsetzung wie Evaluationen des Musters Reaktion bedacht, unterscheiden sie sich von diesen jedoch durch die bereits vor Evaluationsbeginn feststehende und von den Ergebnissen unabhängige Entscheidungsfindung. Alibi-Typen kennzeichnen weder Erkenntnis-

noch Handlungs- oder Verwendungsinteresse. Sie haben lediglich Bestätigungs- oder Abschlusscharakter, gefragt ist deshalb Zweckrationalität, jedoch keine Sinnrationalität, und es gilt die Devise: Evaluation durchführen und sich nicht beirren lassen. Deshalb steht auch die Frage nicht zur Diskussion, ob Evaluationen in der Lage sind, den für weitere Entscheidungen relevanten Informationsbedarf zu befriedigen. Evaluationswissen, auch wenn adressatenorientiert aufbereitet, wird kaum weiterverarbeitet, sondern eher ignoriert oder bestenfalls argumentativ zur Rechtfertigung festgelegter Entscheide benutzt. Fundierte methodologische oder technische Evaluatorenarbeit hat ebenso geringe Relevanz wie die Verwendungstauglichkeit der Befunde, obwohl sie in Evaluationen dieses Typs durchweg positiv eingeschätzt wird. Damit wird nur allzu deutlich, dass Rezeption und Gebrauch nicht lediglich davon abhängen, ob Nutzen- und Anwendungseinschätzung positiv ausfallen, sondern weit stärker mit anderen kontextualen Mechanismen verbunden sind. Deshalb gestaltet sich die Arbeit von Evaluatorinnen und Evaluatoren in einer Alibi-Typik als schwierige Aufgabe. Sie sind nicht nur stark auf sich gestellt, sondern können sich auch kaum entfalten und werden darüber hinaus mit strikten Vorgaben und Regeln eingedeckt, die ihre Rolle im Hinblick auf Dissemination und Datennutzung bestimmen. Verständlich also, wenn sich Evaluatorinnen und Evaluatoren ärgern und sich als Opfer irregeleiteten Verhaltens und falscher Vorstellungen empfinden, wenn sie feststellen müssen, dass ihre Arbeit nicht ernst genommen wird, die Befunde lediglich zur Rechtfertigung der erwünschten Praxis dienen und den Einsatz politischer Macht verfolgen, jedoch nicht als Instanzen kritischer Handlungskontrolle und rationaler Handlungssteuerung betrachtet werden. Sie müssen erkennen, dass ihre Evaluationen Informationen liefern, die nicht gebraucht, nicht zur Kenntnis genommen werden und ohne vordergründige Folgen bleiben. Verdeckt bleibt allerdings die wichtige, strategisch-symbolische Nutzung, die in der nachträglichen Legitimierung dessen besteht, was vorgängig aus anderen Gründen entschieden worden ist, ein Vorgang, für den Drerup's Bezeichnung „legitimatorische Legendbildung" (1982:165) zutreffend ist.

5.4 Die Typik der Reaktion

Dieser Typus beinhaltet Evaluationen, die zwar auf die Überprüfung (44%) und Kontrolle (55%) ausgerichtet sind, jedoch eine klare Rezeption und Nutzung der Ergebnisse vorweisen können (100%). Am augenfälligsten ist dabei, dass dies als Reaktion geschieht, bestimmte Beschlüsse oder Maßnahmen infolgedessen direkte Antworten auf Forderungen bestimmter Anspruchsgruppen darstellen. Die Evaluation, durchgeführt von reputierten Evaluatorinnen und Evaluatoren (100%), hat relativ hohe bildungspolitische Bedeutung (78%), muss aber auch mit ausgeprägtem Widerstand einzelner Akteurgruppen kämpfen (78%).

Die Definition dieser Typik gründet in der Schnittstelle zwischen direkter Rezeption und der Vergleichsdimension Überprüfung/Kontrolle. Obwohl die Entscheidungsverantwortlichen die Schlüsselrolle im Umsetzungsprozess übernehmen, sind es die bildungspolitischen Behörden, welche die Entscheide treffen. Das mag ein Grund dafür sein, dass sich teilweise verkürzte Verwendungsinteressen asymmetrischer oder gar komplementärer Art durchsetzen können, die – ohne Rücksicht auf

die Botschaft der Evaluation – technologische, praktische, aber auch ideologische Ansprüche verfolgen und dem Zweck oder Sinn entgegengesetzte Nutzungsanstrengungen nach sich ziehen, ohne Rücksichtnahme auf das persönliche Wohlbefinden einzelner Individuen oder Gruppen. In dieser Hinsicht unterscheidet sich die Reaktion von der Innovation, die eher auf diskursiv ausgehandelte Nutzungswege und Lerneffekte setzt. Dort, wo die Reaktions-Typik von vorwiegend technologischen Verwendungsinteressen geprägt ist, gelten solche Erkenntnisse als verwendungsrelevant, die objektiv sind und strengen wissenschaftlichen Kriterien genügen. Von Evaluatorinnen und Evaluatoren wird deshalb erwartet, dass sie die letzten Sicherheiten pädagogischer Argumentation liefern, um die Evaluation mit dem Siegel wissenschaftlicher Redlichkeit auszeichnen zu können.

Im Mittelpunkt der Reaktions-Typik stehen deshalb Versachlichungshoffnungen und ein hohes Handlungsinteresse der Entscheidungsverantwortlichen. Folglich besteht großes Interesse an Evaluationsergebnissen, die spezifisch methodisch abgesichert und bestätigt sind und deshalb als gesicherte Grundlagen gegenüber Mutmaßungen, spekulativem Vorwissen und dem Widerstand der Feldakteure eingesetzt und der bildungspolitischen Bedeutung der Evaluation gerecht werden können. Deshalb wird im Reaktions-Typus die Umsetzung der Ergebnisse, das Praktisch-Werden der Befunde, prioritär behandelt, der Erkenntniszuwachs tritt dabei als willkommenes Nebenprodukt auf. Ähnlich dem Blockade-Typus, jedoch in deutlicher Abgrenzung zur Innovation, basieren Umsetzungshandlungen auf der Einschätzung der Verwendbarkeit in Anlehnung an politisch gefärbte Konstellationen. Positive Ergebnisse bringen dort Bestätigung und Entlastung, wo sie wie erwartet ausfallen, negative Erfolgsbilanzen führen jedoch zu Kritik, auf die mit schnellem Entscheiden reagiert wird. Dass auch Evaluationen mit der Zielsetzung Überprüfung/Kontrolle direkte Rezeption und Verwendung erzeugen, widerspricht dem in der Evaluationsliteratur postulierten üblichen Bild. Allerdings verkörpert die Reaktion nicht das Beispiel einer reibungslosen Transferevaluation, die Wissen gradlinig in Umsetzung befördert, sondern ein von erheblichen Unruhen begleitetes Unternehmen, das Folgeaktivitäten als ‚Notmaßnahmen' geradezu provoziert.

6. Bilanz oder: Was bleibt?

Im Ergebnis zeigen die Befunde der Feldstudie ein unangenehmes Maß an Zwiespältigkeit: Zwar kann Evaluation eindeutig vom möglichen Vorwurf der Folgenlosigkeit freigesprochen werden, denn es sind ebenso viele Evaluationen zu eruieren, die nachfolgende bildungspolitische Entscheide wesentlich beeinflusst haben, wie auch Evaluationen, die auf anschließende Veränderungs- und Optimierungsmaßnahmen keinen Einfluss hatten. Evaluation ist somit keine unhaltbare Wunschvorstellung der Evaluationsforschenden und auch keine *grande illusion,* sondern zumindest in der Hälfte der Fälle pädagogische Praxis. Eine differenzierte Betrachtung der Nutzungsperspektiven und -verläufe macht aber klar, dass das Verwendungsinteresse der Aufraggeber- und Anwendersysteme gerade in diesen Fällen unterschiedlich ist, je nachdem, ob die Evaluation eher zur Überprüfung/Kontrolle, zur Legitimation, zur Innovation/Entwicklung oder lediglich zur Bestätigung

durchgeführt wird. Umgekehrt ist der Befund, dass die Hälfte der erfassten Evaluationen keine direkten Folgen verzeichnen kann, noch kein Beweisstück für die Wirkungslosigkeit vieler Evaluationen an sich, sondern – und dies macht die ganze Angelegenheit noch komplizierter – eine Frage des Rahmenkonzepts von Evaluationsnutzung. Gerade die unterschiedlichen Standpunkte, welche vorab in der amerikanischen *Research on evaluation utilization* vertreten werden – ob lediglich direkte, instrumentelle Nutzung als Verwendung gelten soll oder auch die konzeptualisierende oder symbolische Nutzung – verweisen auf die unterschiedlichen Einschätzungen der Wirkungen. Dazu kommt, dass auch die Ansätze und Wege bei der direkten Nutzung so vielschichtig und diffus sind, dass sie kaum klare Antworten zulassen und somit auch kaum als Beweise *gegen* das Scheitern von Evaluation und *für* ihre Legitimation herangezogen werden können. Folgen von Evaluation, so die bilanzierende Erkenntnis, sind zu komplex, als dass sie sich im Lichte „ihrer Nützlichkeit für nachfolgende Entscheidungen" (Rossi/Freeman/Lipsey 1999: 431) eindeutig bewerten lassen könnten.

Die Erfahrungen mit diesen empirischen Ergebnissen verändern die Erkenntnisse über Art, Umfang und Wege der Evaluationsnutzung. Im Ganzen gesehen legen sie nahe, dass direkte Rezeption und Nutzung zwar deutlich häufiger festgestellt werden können als in der einschlägigen theoretischen Literatur behauptet, jedoch nicht lediglich als gradlinige Transferhandlungen vonstatten gehen, sondern in größerem Ausmaß auch über differenzierte Transformationshandlungen theoretischer und praktischer Art. Direkte Nutzung von Evaluationsinformation ist kein unhaltbarer Mythos, keine Wunschvorstellung der Forschung, sondern pädagogische Realität. Sie findet in den beiden Typiken Innovation und Reaktion ihren Ausdruck und demontiert damit Annahmen über die ausbleibende direkte Verwendung von Evaluationswissen. Zugleich geben diese beiden Typen zu erkennen, dass auch direkte Nutzung vielschichtig verläuft und die Anteile, welche nicht im intendierten Schema der „rationalen Chronologie" (Weiss 1981: 28) ablaufen, sondern verschlungenere Wege gehen, beträchtlich sind. Evaluationsfolgen sind somit keinesfalls über jeden Zweifel erhaben.

Gesamthaft lassen die problematischen Befunde der hier präsentierten empirischen Studie und die Erkenntnisse über die verschlungenen Wege des Praktischwerdens des generierten Wissens die Wirksamkeit von Evaluation in einem problematischeren Licht erscheinen als dies in Anbetracht der aktuellen bildungspolitischen Großwetterlage wünschbar wäre. Ohne allerdings ihre kulturelle Bedeutung als *grande idée* vorschnell aufs Spiel zu setzen und sich in die Reihe derer einzuordnen, die Evaluation als *grande illusion* abtun, ist zumindest dann dem Glauben an ihre Kraft mit einem gewissen Maß an Skepsis zu begegnen, wenn sie unhinterfragt und als moralische Instanz eines bildungspolitisch schlechten Gewissens zur Lenkung der Qualität der Bildungssysteme eingesetzt wird.

Sollten Evaluationen angesichts der skizzierten Befunde nicht eher unterbleiben, wenn Entscheidungen großmehrheitlich auf der selektiven Wahrnehmung von Evaluationsbefunden basieren und diese häufig nicht mehr als eine Stimulation des Wissens der Abnehmer bewirken können, jedoch damit rechnen müssen, nicht gebraucht, uminterpretiert, relativiert und trivialisiert zu werden? Wäre es eine vordringliche Aufgabe der Evaluationsforschung, vor Evaluation zu warnen? Eine derartige Konsequenz wäre nun allerdings doch eine zu krasse Folgerung aus den bis-

herigen Ausführungen. Die empirischen Befunde dieses Aufsatzes legen vielmehr nahe, die harmonische Lösbarkeit der Nutzungsfrage als skeptisches Unterfangen zu akzeptieren und sich unter Verzicht auf positive Meinungspflege und kritiklose Akzeptanz im Prokrustesbett zurechtzufinden. Deshalb gilt es, die Legitimation von Evaluation auf das Postulat der Verwendbarkeit auszurichten, d.h. auf die Hauptaufgabe, in geschickter Art und Weise Verwendung zu provozieren, aber mit der Einsicht, dass sie nie oder nur in seltenen Fällen den Vorstellungen der Evaluationsforschung entsprechen kann, also inadäquat bleiben muss. Das Fatale daran ist, dass diese Erkenntnis immer sowohl Belastung wie auch Entlastung sein kann und einer erheblichen Ideologieanfälligkeit unterliegt. Wenn Nutzung zwar provoziert werden soll und teilweise auch tatsächlich nachgewiesen werden kann, jedoch nur ausnahmsweise den Vorstellungen der Evaluationsforschung entspricht, dann haben Evaluatorinnen und Evaluatoren Verantwortung für diese provozierenden Aktivitäten zu tragen. Gleichzeitig gilt auch, Evaluation für eine Reformierung von Maßstäben der Nutzenerwartung frei zu machen. Um dieses Dilemma lösen zu können, sind keine destruierenden Rundumschläge oder enthusiastischen Überzeugungsarbeiten gefragt, die sich etwa in Extrembekenntnissen artikulieren, eine evaluatorenbesetzte Gesellschaft solle die Welt demokratisieren und aus ihr einen besseren und weiseren, weniger von Konflikten besetzten Platz machen (Patton 2000). Eher gefragt scheint ein Plädoyer für mehr Distanz zu Evaluation, für ihren reflektierteren Einsatz, der sie zwar zu einem knapperen Gut werden lässt und zwangsläufig zu einer Redimensionierung des Evaluationskults führt, ihr aber vielleicht die Chance eröffnet, gezieltere Wirksamkeit als Bildungsprozess zu entfalten.

7. Literatur

Alkin, M.C. (1980): Naturalistic study of evaluation utilization. In: New Directions for Program Evaluation (Utilization of Evaluative Information), 5, S. 19-28.

Alkin, M.C./Coyle, K. (1988): Thoughts on evaluation utilization, misutilization and non-utilization. In: Studies in Educational Evaluation, 14, S. 331-340.

Altrichter, H./Posch, P. (1997): Mikropolitik der Schulentwicklung. Förderliche und hemmende Bedingungen für Innovationen in der Schule. Innsbruck: StudienVerlag.

Backhaus, K. (1996): Multivariate Analysemethoden – eine anwendungsorientierte Einführung. Berlin: Springer.

Beck, U./Bonß, W. (1989a): Verwissenschaftlichung ohne Aufklärung? In: Beck, U./Bonß, W. (Hg.): Weder Sozialtechnologie noch Aufklärung? Analysen zur Verwendung sozialwissenschaftlichen Wissens. Frankfurt a. M.: Suhrkamp, S. 7-45.

Beck, U./Bonß, W. (1989b): Weder Sozialtechnologie noch Aufklärung? Analysen zur Verwendung sozialwissenschaftlichen Wissens. Frankfurt a. M.: Suhrkamp.

Beck, U./Lau, Ch. (1983): Die „Verwendungstauglichkeit“ sozialwissenschaftlicher Theorien: Das Beispiel der Bildungs- und Arbeitsmarktforschung. In: Beck, U. (Hg.): Soziologie und Praxis. Erfahrungen, Konflikte, Perspektiven. Sonderband 1 der Zeitschrift Soziale Welt. Göttingen, S. 369-402.

Benner, D. (2002): Die Struktur der Allgemeinbildung im Kerncurriculum moderner Bildungssysteme. Ein Vorschlag zur bildungspolitischen Rahmung von PISA. In: Zeitschrift für Pädagogik, 1, S. 68-90.

Beywl, W. (1988): Zur Weiterentwicklung der Evaluationsmethodologie. Grundlegung, Konzeption und Anwendung eines Modells der responsiven Evaluation. Bern: Lang (Reprint 1999).

Burkard, C./Eikenbusch, G. (2000): Praxishandbuch Evaluation in der Schule. Berlin: Cornelsen.
Cousins, J.B./Leithwood, K.A. (1986): Current empirical research on evaluation utilization. In: Review of Educational Research, 56 (3), S. 331-364.
Cousins, J.B./Leithwood, K.A. (1993): Enhancing knowledge utilization as a strategy for school improvement.In: Knowledge: Creation, Diffusion, Utilization, 14 (3), S. 305-33.
Daillak, R.H. (1982): What is evaluation utilization? In: Studies in Educational Evaluation, 8, S. 157-162.
Dickey, B. (1981): Utilization of evaluation of small-scale educational projects. In: Educational Evaluation and Policy Analysis, 2 (6), S. 65-77.
Drerup, H. (1979): Wissenschaftstheorie und Wissenschaftspraxis. Probleme der Vermittlung zwischen metawissenschaftlichen Forschungsprogrammen und einer Praxis der Spezial-/Erziehungswissenschaft. Grundmann: Bonn.
Drerup, H. (1982): Anwendungsprobleme in der Evaluation. In: Bildungsforschung/Bildungspraxis, 4 (2), S. 154-170.
Drerup, H. (1989): Probleme außerwissenschaftlicher Verwendbarkeit von Erziehungswissenschaft. Zum Einfluss von Erziehungswissenschaft im politisch-administrativen Bereich. In: König, E./Zedler, P. (Hg.): Rezeption und Verwendung erziehungswissenschaftlichen Wissens in pädagogischen Handlungs- und Entscheidungsfeldern. Weinheim: Deutscher Studienverlag, S. 143-166.
Drerup, H./Terhart, E. (1990): Erkenntnis und Gestaltung. Vom Nutzen erziehungswissenschaftlicher Forschung in praktischen Verwendungskontexten. Weinheim: Deutscher Studienverlag.
Dubs, R. (2000): Teilautonomie der Schulen: Annahmen, Begriffe, Probleme, Perspektiven. Paderborn: Paderborner Universitätsreden.
Evers, A./Nowotny, H. (1989): Über den Umgang mit Unsicherheit. Anmerkungen zur Verwendung sozialwissenschaftlichen Wissens. In: Beck, U./Bonß, W. (Hg.): Weder Sozialtechnologie noch Aufklärung? Analysen zur Verwendung sozialwissenschaftlichen Wissens. Frankfurt a. M.: Suhrkamp, S. 355-383.
Gather Thurler, M./Huberman, M. (1993): Umsetzung wissenschaftlicher Forschung in die Praxis: unter welchen Bedingungen sind Forscher dazu fähig und bereit? In: Bildungsforschung und Bildungspraxis, 3, S. 309-325.
Geertz, C. (1999): Dichte Beschreibung. Beiträge zum Verstehen kultureller Systeme. Frankfurt a. M.: Suhrkamp.
Grawe, K., Donati, R./Bernauer, F. (1993): Psychotherapie im Wandel. Von der Konfession zur Profession. Göttingen: Hogrefe.
Greene, J.C. (1987): Stakeholder partizipation in evaluation design: is it worth the effort? In: Evaluation and Program Planning, 10, S. 379-394.
Guba, E.G./Lincoln, Y.S. (1981): Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approach. San Francisco: Jossey-Bass.
Hagemeister, V. (1999): Was wurde bei TIMMS erhoben? Über die empirische Basis einer aufregenden Studie. In: Die Deutsche Schule, 2, S. 160-177.
Hagemeister, V. (2000a): Irrwege und Wege zur „Testkultur“. In: Die Deutsche Schule, 1, S. 322-328.
Hagemeister, V. (2000b): Die TIMSS-Leistungen bleiben zweifelhaft. In: Die Deutsche Schule, 3, S. 480-490.
Heller, F. (1986): The use and abuse of social science. London: Sage.
Hellstern, G.-M./Wollmann, H. (1983): Evaluierungsforschung. Ansätze und Methoden – dargestellt am Beispiel des Städtebaus. Stuttgart: Birkhäuser.
Helmke, A. (2000): Von der externen Leistungsevaluation zur Verbesserung des Lehrens und Lernens. In: Trier, U.P. (Hg.): Bildungswirksamkeit zwischen Forschung und Politik. Chur: Rüegger, S. 135-164.
Kelle, U./Kluge, S. (1999): Vom Einzelfall zum Typus. Opladen: Leske + Budrich.
Kluge, S. (2000): Empirisch begründete Typenbildung in der qualitativen Sozialforschung. Forum Qualitative Sozialforschung 1 (1) [On-line]. Available: http: www.qualitative-research.net/fqs.

Kordes, H. (1983): Evaluation in Curriculumprozessen. In: Hameyer, U./Frey, K./Kraft, H. (Hg.): Handbuch der Curriculumforschung. Weinheim/Basel: Beltz.
Krapp, A./Weidenmann, B. (Hg.) (2001): Pädagogische Psychologie. Ein Lehrbuch. Weinheim: Beltz, Psychologie Verlags Union.
Kuckartz, U. (1999): Computergestützte Analyse qualitativer Daten. Eine Einführung in Methoden und Arbeitstechniken. Opladen: Westdeutscher Verlag.
Lange, H. (1999): Qualitätssicherung in Schulen. In: Die Deutsche Schule, 1, S. 144-159.
Luhmann, N./Schorr, K.E. (1982): Zwischen Technologie und Selbstreferenz. Fragen an die Pädagogik. Frankfurt a. M.: Suhrkamp.
Nuthmann, R. (1983): Qualifikationsforschung und Bildungspolitik – Entwicklungen und Perspektiven. In: Zeitschrift für Sozialisationsforschung und Erziehungssoziologie, 3 (2), S. 175-188.
Oelkers, J. (1997): Effizienz und Evaluation in der Lehrerausbildung. In: Beiträge zur Lehrerbildung, 1, S. 15-25.
Oelkers, J. (2001): Bildung, Bildungsforschung und Bildungspolitik. Vortrag am Züricher Festival des Wissens, 10. Mai.
Patton, M.Q. (1986, 1997): Utilization-focused evaluation. Thousand Oaks, London, New Dehli: Sage.
Patton, M.Q. (2000): A vision of evaluation that strengthens democracy. Referat an der 4. Jahreskonferenz der Europäischen Evaluationsgesellschaft in Lausanne, 14. Oktober.
Pekrun, R. (2002): Vergleichende Evaluationsstudien zu Schülerleistungen: Konsequenzen für zukünftige Bildungsforschung. In: Zeitschrift für Pädagogik, 1, S. 111-118.
Rossi, P.H./Freeman, H.E./Lipsey, M.W. (1999): Evaluation: a systematic approach, 6th edition. Thousand Oaks, London, New Dehli: Sage.
Rost, D.H. (2001): Handwörterbuch Pädagogische Psychologie. Weinheim: Beltz, Psychologie Verlags Union.
Shulha, L.M./Cousins, J.B. (1997): Evaluation use: Theory, research, and practice since 1986. In: Evaluation Practice, 18 (3), S. 195-208.
Siegel, K./Tuckel, P. (1985): The utilization of evaluation research. In: Evaluation Review, 9 (3), S. 307-328.
Terhart, E. (2002): Wie können die Ergebnisse von vergleichenden Leistungsstudien systematisch zur Qualitätsverbesserung in Schulen genutzt werden? In: Zeitschrift für Pädagogik, 1, S. 91-110.
Tillmann, K.-J./Vollstädt, W. (2001) (Hg.): Politikberatung durch Bildungsforschung. Das Beispiel: Schulentwicklung Hamburg. Opladen: Leske + Budrich.
Von der Groeben, A. (2000): Wo liegen die Wurzeln von Schulqualität? Eine Antwort auf Hermann Lange. In: Die Deutsche Schule, 2, S. 339-354.
Von der Groeben, A. (2002): Nicht in Maßnahmen stecken bleiben. Ein Plädoyer für eine radikale Frage nach PISA. Pädagogik, 4, S. 38-42.
Weber, U. (1994): Regulation und Wissen: Implikationen neuerer Ergebnisse der Verwendungsforschung für eine Theorie der Regulation. Frankfurt a. M., Bern: Lang.
Weinert, F.E. (2002): Perspektiven der Schulleistung – mehrperspektifisch betrachtet. In: Weinert, F.E (Hg.): Leistungsmessung in Schulen. Weinheim, Basel: Beltz, (S. 353-365).
Weiss, C.H. (1972a): Evaluation research: Methods of assessing program effectiveness. Englewood Cliffs, NJ: Prentice-Hall.
Weiss, C.H. (Hg.). (1972b): Evaluation action programs: Readings in social action and education. Boston: Allyn and Bacon.
Weiss, C.H. (1972c): Utilization of evaluation. Toward comparative study. In: Weiss, C.H. (Hg.): Evaluation action programs: readings in social action and education. Boston: Allyn and Bacon, S. 318-326.
Weiss, C.H. (1981): Measuring the use of evaluation. In: Ciarolo, J.A. (Hg.): Utilizing evaluation: Concepts and measurement techniques. Beverly Hills: Sage, S.. 17-33.
Weiss, C.H. (1982): Measuring the use of evaluation. In: Evaluation Studies Review Annual, 7, S. 129-145.

Weiss, C.H. (1998): Have we learnt anything new about the use of evaluation? In: American Journal of Evaluation, 19 (1), S. 21-23.
Weiss, C.H./Bucuvalas, M.J. (1980): Social science research and decision-making. New York: Columbia University Press.
Witzel, A. (1982): Verfahren der qualitativen Sozialforschung. Frankfurt a. M.: Campus.
Wottawa, H./Thierau, H. (1998). Lehrbuch Evaluation. Bern: Huber.