Evaluationsforschung in der Psychologie

Helfried Moosbrugger[1] und Karl Schweizer[2]

1. Einleitung

In Abwandlung eines bekannten Zitats von Hermann Ebinghaus (1908) über die Psychologie kann festgestellt werden, dass die psychologische Evaluationsforschung „eine lange Vergangenheit, doch nur eine kurze Geschichte“ hat. Schon lange bevor der Evaluationsbegriff in die wissenschaftliche und öffentliche Diskussion Eingang gefunden hatte, wurden bereits Objekte, Maßnahmen oder Ideen bewertet. Der Beginn der modernen Evaluation wird jedoch erst in der Mitte dieses Jahrhunderts datiert (vgl. Cook & Matt 1990), als in den USA die Einführung umfangreicher und teurer sozialer Programme anstand, deren Nutzen im Sinne einer ethisch-moralischen Verantwortung unter Beweis zu stellen war (vgl. Wottawa & Thierau 1998: 14f.). Aus Anlass solcher Programme wurde damit begonnen, geeignete wissenschaftliche Methoden zu entwickeln. Diese Entwicklung war allerdings zunächst auf die USA beschränkt und hat erst Ende der siebziger Jahre auf den deutschsprachigen Raum übergegriffen (vgl. z.B. Lange 1983; Wittmann 1985; Wottawa & Thierau 1998: 58). Mit der Einführung eines Studien- und Prüfungsfachs „Evaluation und Forschungsmethoden“ im Diplomstudiengang Psychologie und eines entsprechenden Curriculums für die Ausbildung (Moosbrugger, Rost & Schermelleh-Engel 1999) vermochte sich die Evaluationsforschung in der Psychologie endgültig zu etablieren.

1.1 Zum Begriff der Evaluation

Mit dem Begriff Evaluation verbinden sich eine Reihe unterschiedlicher Vorstellungen. Ganz allgemein steht Evaluation für die Festsetzung des Wertes einer Sache (Moosbrugger 1994; Wottawa & Thierau 1998: 14). In der Psychologie bewirkt die inhaltliche Bestimmung des Faches darüber hinaus die Zentrierung auf psychologische Maßnahmen als Sache. Es handelt sich um Maßnahmen mit einer Wirkung,

1 Universität Frankfurt,
2 Deutsches Institut für Internationale Pädagogische Forschung Frankfurt

die Menschen einen Nutzen in Aussicht stellt. Die Evaluation soll eine Überprüfung der Maßnahmen im Sinne einer Bewertung erbringen. So wird unter Evaluation etwa die Überprüfung von psychologisch-therapeutischen Interventionsmaßnahmen oder auch andere Wirksamkeitsprüfungen verstanden (Hager, Patry & Brezing 2000: 1). Im Vordergrund steht dabei der Nachweis der Wirksamkeit dieser Maßnahme; eine Erklärung für die Wirksamkeit braucht die Evaluation dagegen nicht zu erbringen.

Dieser Anspruch des Nachweises von Wirksamkeit kommt auch in der besonderen ethisch-moralischen Komponente von Evaluation zum Ausdruck. Durch die Evaluation soll der Nutzen einer Maßnahme für Menschen sichergestellt werden, da sich mit solchen Maßnahmen mitunter erhebliche Konsequenzen für die Menschen verbinden können (Hager, Patry & Brezing 2000: 1; Wottawa & Thierau 1998: 3ff.), so dass zwischen Evaluationsmaßnahmen und den betroffenen Menschen ein viel direkterer Bezug besteht als in vielen anderen Bereichen der Humanwissenschaften. Diese ethisch-moralische Komponente verpflichte dazu, dass besonders hohe Anforderungen an die der Evaluation immanente Bewertung gestellt werden. Im Hinblick auf dieses Ziel wird im Rahmen der Evaluation eine Handlungsoptimierung gefordert (Hager, Patry & Brezing 2000), und zwar sowohl bezüglich einer summativen Evaluation, die sich auf das Ergebnis der Maßnahme richtet, wie auch bezüglich einer formativen Evaluation, die im laufenden Prozess auf die Anwendung der Maßnahme Einfluss nimmt (Scriven 1980). Die Bewertung von Maßnahmen sollte jedenfalls höchsten Ansprüchen genügen. In diesem Sinne erfolgt in zunehmendem Maße die Auseinandersetzung mit Gütemerkmalen, Kriterien und Standards für Evaluation (vgl. z.B. Deutsche Gesellschaft für Evaluation 2002; Hager & Hasselhorn 2000; Hager, Patry & Brezing 2000; Joint Commitee on Standards for Educational Evaluation 1994; Rost 2000; Wottawa & Thierau 1998). Durch solche Gütemerkmale, Kriterien und Standards soll gewährleistet werden, dass die Wahrscheinlichkeit von fehlerhaften Entscheidungen und Empfehlungen aufgrund einer Evaluation minimiert wird.

Die Kombination von wissenschaftlicher Methodik und Evaluation erlaubt es, zwischen Evaluation und Evaluationsforschung zu unterscheiden (vgl. Moosbrugger 1994). Unter Evaluation (ohne -forschung) wird in der Regel ein Bewertungsprozess verstanden, in dem der Wert eines Produkts, einer Maßnahme oder eines Programms beurteilt, ggf. auch nur behauptet wird. Die Evaluationsforschung hingegen steht für die Optimierung der Überprüfung von Maßnahmen (Hager, Patry & Brezing 2000), bei der wissenschaftliche, datengestützte Verfahren zur empirischen Untermauerung der Beurteilung Verwendung finden. Somit wird die Evaluationsforschung mit der systematischen Anwendung von Prozeduren der empirischen Sozialforschung assoziiert (Rossi & Freeman 1989). Diese Unterscheidung von Evaluation und Evaluationsforschung schafft die Möglichkeit, die alltagsprachliche Bewertung von der wissenschaftlichen Bewertung zu trennen und befördert damit die Eindeutigkeit in der wissenschaftlichen Kommunikation. Sie konnte sich bisher allerdings noch nicht einheitlich im Sprachgebrauch der Wissenschaftler durchsetzen.

In der Psychologie kommt Evaluation vor allem in den folgenden vier inhaltlichen Bereichen zum Einsatz: Evaluation sozialer Programme, Therapieevaluation, Evaluation von Maßnahmen und Innovationen im Kontext der Arbeits- und Organi-

sationspsychologie und Evaluation im pädagogischen Bereich (Rost 2000). Hinzu kommt als fünfter Bereich die Lehrevaluation, welche von der Evaluation im pädagogischen Bereich insoweit abgegrenzt werden kann, als im Rahmen der Lehrevaluation Psychologen nicht nur als Evaluatoren, sondern z.T. auch als Evaluierte auftreten, wodurch sich eine Reihe zusätzlicher Probleme und Fragestellungen ergeben (s.u.). Natürlich können auch differenzierte Klassifikationen für den Anwendungsbereich vorgenommen werden (vgl. Wottawa & Thierau 1998).

2. Allgemeine Entwicklungen in der psychologischen Evaluationsforschung

2.1 Wachsender wissenschaftlicher Anspruch

In der Evaluationsforschung kann ein wachsender wissenschaftlicher Anspruch konstatiert werden. War die Evaluationsforschung in den Anfängen noch in hohem Maße Improvisation, so ist sie heute in vielerlei Hinsicht durch wissenschaftliche Übereinkünfte, Kriterien und Standards gesteuert; die Fortsetzung dieses Trends wird auch die nähere Zukunft der Evaluationsforschung bestimmen. Im Hinblick auf viele Fragestellungen fehlte es zunächst an geeigneten Messinstrumenten. Konsequenterweise bestand eine wichtige Aufgabe in der Erarbeitung von Richtlinien zur Herstellung von ad-hoc Messinstrumenten (vgl. Henerson, Morris & Fitz-Gibbon 1987; Morris, Fitz-Gibbon & Lindheim 1987). Solche ad-hoc Messinstrumente waren also in vielen Fällen zunächst nicht das Ergebnis eines systematischen Konstruktionsprozesses. In der Zwischenzeit sind geeignete Messinstrumente in zunehmendem Maße auf Vorrat bereitgestellt worden oder können aus früheren Evaluationsprojekten übernommen und verbessert werden wie etwa im Bereich der Lehrevaluation (s.u.).

Mit dem Aufkommen der Evaluationsforschung als Auftragsforschung hat sich eine neue Problematik ergeben, die darin besteht, dass die Evaluationsforschung in der Regel nicht die Freiheit der Forschung für sich in Anspruch nehmen kann, sondern an einen Auftraggeber gebunden ist. Wegen dieser Verantwortlichkeit gegenüber einem Auftraggeber müssen sich Forscher bezüglich der Akzeptanz ihrer Arbeit mit dem Auftraggeber und dessen Interessen auseinandersetzen (vgl. Wottawa & Thierau 1998: 23ff.). Dies stellt aber auch eine neue Herausforderung dar, welche zur Unterscheidung verschiedener Herangehensmöglichkeiten geführt hat, die mit Evaluationsansatz, experimentellem Ansatz, zielgerichtetem Ansatz, entscheidungsorientiertem Ansatz und responsivem Ansatz bezeichnet werden (Stecher & Davis 1987: 22ff.). Um die Akzeptanz für Evaluation zu erhöhen, wird die Auseinandersetzung mit den Beteiligten und die Rücksichtnahme auf deren Interessen dringend empfohlen, was aber in der Regel zu einer eingeschränkten Realisierbarkeit des wissenschaftlichen Anspruchs führt und insoweit beklagt wird (Spiel 2001). Es ist allerdings zu erwarten, dass dieses Dilemma zwischen Akzeptanz und methodischen Anspruch sich in absehbarer Zeit verringern wird, da Einsicht in Sinn und Notwendigkeit von Evaluation in zunehmendem Maße bei Auftraggebern vorausgesetzt werden kann und das Selbstverständnis des Evaluators durch die fortlau-

fende Vergrößerung des methodischen Inventars gestärkt wird. Durch die Entwicklung und Erforschung geeigneter Methoden der Evaluation wird das methodische Inventar in Form von Richtlinien für die Planung durch feste Ablaufpläne, durch Kriterien und Standards für das Vorgehen im Einzelnen und durch die Antizipation vieler Eventualitäten immer mehr vergrößert. Umgekehrt kann festgestellt werden, dass sich der Spielraum für unangemessene Anpassung an die Wünsche von Auftraggebern im Sinne von Abstrichen hinsichtlich der Qualität einer Evaluationsstudie immer mehr verringert.

2.2 Unterscheidung von Evaluationsforschung und Grundlagenforschung

Aufgrund dieser Entwicklungen präsentiert sich die Evaluationsforschung in zunehmendem Maße als neues eigenständiges Forschungsparadigma neben dem primärem Forschungsparadigma der Grundlagenforschung. Der Vergleich zwischen Grundlagenforschung und Evaluationsforschung lässt einige Unterschiedlichkeiten erkennen (vgl. Rost 2000). Während die Grundlagenforschung die Untersuchung theoretisch begründeter Fragestellungen zum Ziel hat, steht bei der Evaluationsforschung die Untersuchung von Fragen der Wirksamkeit im Vordergrund. Bei der Grundlagenforschung ist ein theoretisches Problem Gegenstand der Untersuchung, bei der Evaluationsforschung hingegen eine Intervention, worunter so unterschiedliche Dinge wie ein Trainingsprogramm, ein Curriculum oder auch eine Werbemaßnahme verstanden werden können. Bei der Grundlagenforschung geht es meist um die Untersuchung einer aufgrund klarer theoretischer Vorgaben punktgenauen Fragestellung, wohingegen es sich bei der Evaluationsforschung um einen nicht in allen Einzelheiten theoretisch durchdrungenen und daher komplexen Gegenstand handelt, woraus sich als Konsequenz ergibt, dass die Kontrolle von Störbedingungen durch Konstanthaltung oder Randomisierung in der Grundlagenforschung meist vorgenommen werden kann, während sie in der Evaluationsforschung nur eingeschränkt oder sogar überhaupt nicht realisierbar ist. Schließlich muss auch auf die unterschiedlichen Ergebniserwartungen hingewiesen werden. Der Grundlagenforschung kommt es auf den Nachweis der Unterschiedlichkeit im Sinne eines Treatmenteffekts an. Im Gegensatz dazu trachtet die Evaluationsforschung gewöhnlich nach dem Nachweis der Effektivität, die sich nicht einfach auf Signifikanz im Sinne eines Treatmenteffekts beschränken kann, sondern darüber hinaus den Nachweis der Relevanz dieses Effekts erfordert (z.B. Bortz & Döring 2002; Hager, Patry & Brezing 2000). Aus der Perspektive charakteristischer Phasen in der Evaluations- und Grundlagenforschung lassen sich als Phasen großer Ähnlichkeit insbesondere die Phasen der Konzeptualisierung, der Implementation und der Wirkungsforschung identifizieren (Rossi, Freeman & Hofmann 1988). Solche Ähnlichkeitsfeststellungen sind insoweit nicht überraschend, als die Methodik der Evaluationsforschung sich aus der Methodik der Grundlagenforschung entwickelt hat. Für den weiteren Erfolg der Evaluationsforschung erscheint es allerdings sehr wichtig, dass ihre weitere Entwicklung auch in methodischen Belangen einen eigenständigen Weg nimmt, um den spezifischen Problemstellungen in besonderem Maße gerecht werden zu können.

Der Unterschied zwischen der Grundlagen- und der Evaluationsforschung kommt auch in der Unterscheidung von nomologischen Aussagen und technologischen Aussagen zum Ausdruck (Patry & Perrez 2000). Durch diese Systematisierung auf der theoretischen Ebene wird die Eigenständigkeit der Evaluationsforschung gegenüber der Grundlagenforschung unterstrichen. Während in der Grundlagenforschung nur Anforderungen an die Theorie gestellt und nomologische Aussagen erwartet werden, sind aus der Perspektive der Evaluationsforschung die von Bunge (1967: 132ff.) konzipierten technologischen Aussagen von besonderem Interesse. Patry und Perrez (2000) ordnen den technologischen Aussagen neben der Theorie eine deutliche Verbindung zur Praxis zu. Die Evaluationsforschung muss also den Ansprüchen von Theorie und Praxis genügen, was weitergehende Überlegungen und Maßnahmen erforderlich macht. So ist es für die Praxis nicht nur wichtig zu wissen, ob eine bestimmte Interventionsmaßnahme wirksam ist im Sinne der Erbringung eines Mindesteffekts. Für die Praxis ist es darüber hinaus von Relevanz zu erfahren, wie die Interventionsmaßnahme im Vergleich zu anderen alternativen Interventionsmaßnahmen zu bewerten ist (vgl. Hager 2000; Hager, Patry & Brezing 2000). Diese besonderen Ansprüche an die Evaluationsforschung untermauern geradezu die Forderung nach Eigenständigkeit.

2.3 Zunehmende Differenzierung verschiedener Evaluationsarten

Bei der systematischen Beschreibung verschiedener Arten von Evaluation kann ein zunehmender Grad an Differenzierung von Evaluation beobachtet werden. Zunächst wurde von Scriven (1967, 1972, 1991) die wegweisende Unterscheidung zwischen formativer und summativer Evaluation eingeführt. Aufbauend auf weitergehenden Differenzierungen von Rossi, Freeman und Hofmann (1988), Rossi und Freeman (1993) und Mittag und Jerusalem (1997) unterscheiden Mittag und Hager (2000) zwischen fünf Arten von Evaluation für Interventionsprogramme, nämlich zwischen (1) der Evaluation der Programmkonzeption, (2) der formativen Evaluation, (3) der Evaluation der Programmdurchführung, (4) der Evaluation der Programmwirksamkeit und (5) der Evaluation der Programmeffizienz. Für jede dieser Evaluationsarten entwerfen sie ein Konzept, das zentrale Aufgaben und Arbeitsschritte umfasst. Diese Rahmenkonzeption erlaubt eine frühzeitige Auswahl und Anpassung der Evaluationsart an die Spezifika der jeweiligen Aufgabenstellung.

Ad (1) In Bezug auf die Evaluation der Programmkonzeption gehören dazu die Problembestimmung und Entscheidung über den Bereich der Intervention, die Zielbestimmung, die Konzeption und Gestaltung des Programms, die Auswahl geeigneter diagnostischer Methoden und Verfahren und die Bewertung der Programmkonzeption.

Ad (2) Der formativen Evaluation werden die Bewertung der Programmimplementation, die Bewertung der Zielsetzung des Programms und die Bewertung der Effizienz des Programms als primäre Komponenten zugeordnet.

Ad (3) Für die Evaluation der Programmdurchführung wird zwischen der Kontrolle der Programmausführung und der Prüfung der Programmreichweite unterschieden. Bezüglich der Programmausführung sind Supervision und Überwachung sowie die

Bewertung der Durchführbarkeit unter alltagspraktischen Bedingungen vorgesehen; bezüglich der Programmreichweite sind die Feststellung von Verzerrungen und der Vergleich von Programmteilnehmern und Abbrechern wichtig.

Ad (4) Für die Evaluation der Programmwirksamkeit wird ein vergleichsweise großer Umfang von Arbeitsschritten für notwendig erachtet, nämlich die Steigerung von kurzfristig verfügbaren Kompetenzen, die Verbesserung von langfristig verfügbaren Kompetenzen, Kontrolle des Transfers auf Alltagsbereiche, Merkmale der Programmvermittler und Zielpersonen, Art der Wirkungen des Programms, weitere (potentielle) Fragestellungen und Hypothesen u.a., Prozessevaluation und Metaevaluation bisheriger Evaluationen.

Ad (5) Die Evaluation der Programmeffizienz schließlich erfordert die Erfassung aller Programmkosten und Programmwirkungen, die Bestimmung der Gesamtkosten in Geldeinheiten, die Bestimmung des Gesamtnutzens in Geldeinheiten/die Bestimmung der Gesamtwirkung in Zieleinheiten, die im Einzelfall bestimmt werden müssen, und die Bestimmung der Programmeffizienz.

2.4 Bereitstellung von Standards und Kriterien

Zur Entwicklung einer eigenständigen Evaluationsforschung zählt auch die Bereitstellung von Standards und Kriterien. Auf diese Standards und Kriterien soll von den Evaluatoren im Einzelfall Bezug genommen werden, um Meinungsverschiedenheiten mit Auftraggebern vorzubeugen. Allgemeine Standards für die Evaluationsforschung wurden von der Evaluation Research Society (Rossi 1982) und dem Joint Commitee on Standards for Educational Evaluation (1994) vorgeschlagen. Letzteres umfasst vier Gruppen von Standards, die von Schiffler und Hübner (2000: 142) mit Nutzenstandards, Machbarkeits- oder Durchführbarkeits-Standards, Standards für Anstand und ethisches Vorgehen und Genauigkeits-Standards übersetzt wurden. Die Nutzenstandards fassen die verschiedenen Aspekte des Nutzens der Evaluation für die Beteiligten zusammen. Die Machbarkeits- oder Durchführbarkeits-Standards stehen für die Richtlinien, die bei der Realisierung der Evaluation unter Berücksichtigung der konkreten Randbedingungen beachtet werden sollten. Die Standards für Anstand und ethisches Vorgehen betonen die Einhaltung der ethischen Rechte der Beteiligten. Die Genauigkeits-Standards schließlich sollen einen hinreichenden Grad an Genauigkeit und Ausführlichkeit der Information garantieren, um nützlich zu sein. Diesen Standards kommt auch ein leitender Charakter für die interdisziplinäre Entwicklung der Evaluationsforschung zu.

Eine weitere Sammlung allgemeiner Forderungen liegt von Patry und Perrez (2000: 38) vor, die allerdings primär auf Programme im Sinne von Maßnahmen ausgelegt ist. Patry und Perrez fordern die ethische Legitimation der Ziele und Methoden, die Vereinbarkeit der theoretischen Grundlagen des Programms mit dem rationalen Corpus der sozialwissenschaftlichen Forschung, die Einschätzbarkeit der Wirksamkeit des Programms und eine positive Kosten-Nutzen-Relation. Diese Forderungen an ein Programm können im Rahmen der Evaluationsforschung überprüft werden. Bei den von Hager und Hasselhorn (2000: 81) erarbeiteten Gütekriterien für Interventionsmaßnahmen steht der Nutzen für die von einem Evaluationsprojekt

betroffenen Personen im Mittelpunkt. Sie fordern in Bezug auf solche Maßnahmen ethische Legitimierbarkeit, theoretische Fundierung, empirische Fundierung, den Nachweis des Fehlens von negativen und schädlichen Neben- und Folgewirkungen, die "Bewährung" des Programms in der Praxis, den Nachweis der Verlässlichkeit unter Standard-Randbedingungen, den Nachweis der Robustheit unter veränderten Randbedingungen, die Wirtschaftlichkeit relativ zu den Zielen, die Routinisierbarkeit und Adaptabilität im Sinne der Unabhängigkeit von den konkreten Umständen einer Interventionsmaßnahme, die Akzeptanz des Programms sowie Zufriedenheit mit dem Programm. Im Rahmen der Evaluationsforschung können und sollten diese Punkte einer Überprüfung unterzogen werden.

Zur Sicherung des Wissenschaftlichkeitsstatus werden von Rost (2000: 133ff.) auf der Basis der Gegenüberstellung von Grundlagen- und Evaluationsforschung elf Standardschritte gefordert: Konzeptualisierung der wesentlichen Merkmale der Interventionsmaßnahmen, Entwicklung konzeptbasierter Fragestellungen und Hypothesen, Implementationskontrolle, Untersuchung der Übertragbarkeit von Ergebnissen auf andere Stichproben, Analysen zur Kausalinterpretation der Effekte, Kontrollgruppendesigns, statistische Kontrolle von Drittvariablen, Effektgrößenbestimmung der Wirkung, soziale Sensibilität in Bezug auf die Beteiligten, Transparenz und Nachvollziehbarkeit sowie Wahrung ethischer Maßstäbe.

2.5 Das Evaluationsdesign

Die dargestellten Kriterien und Standards sind sehr allgemein gehalten und bedürfen daher der Anpassung an die jeweilige Situation. Sie stellen jedoch eine nützliche Orientierung für den Evaluator dar, indem sie ihm helfen, langfristig Fehler zu vermeiden, sich effizient auf mögliche Probleme vorzubereiten und den Nutzen zu optimieren. Weitere Entwicklungen in diesem Bereich der Kriterien und Standards, die eine spezifischere Vorbereitung ermöglichen, können erwartet werden. Von den bereits vorliegenden Ausarbeitungen im Sinne konkreter Kriterien und Standards sind besonders diejenigen hervorzuheben, die zur Wahl eines geeigneten Evaluationsdesigns als zentraler Komponente der Validität eines Evaluationsprojekts vorgeschlagen wurden. Diese Ausarbeitungen bauen auf wichtigen grundlegenden Arbeiten zur Validität experimenteller Untersuchungen von Campbell und Stanley (1963, 1973) sowie Cook und Campbell (1979; auch Cook, Campbell & Perachio 1990) auf, in denen Störfaktoren der experimentellen und quasi-experimentellen Forschung benannt und Möglichkeiten zu ihrer Kontrolle reflektieren werden. Dabei wird zwischen Störfaktoren für die interne und die externe Validität unterschieden. Weiterführend kann auf die Überlegungen zur Validität der experimentellen Forschung von Bredenkamp (1980), Hager und Westermann (1983), Hager (1987, 1998), Sarris (1990) und Westermann (1987, 2000) Bezug genommen werden.

Den Ausgangspunkt für eine evaluationsgerechte Validitätskonzeption muss natürlich die auf einer bestimmten Interventionsmaßnahme basierende Erwartung bilden. Gewöhnlich besteht die Erwartung darin, dass die Interventionsmaßnahme eine Verbesserung der Kompetenz im Sinne einer Fertigkeit bzw. eines Skills, bestimmten Anforderungen gerecht werden zu können, nach sich zieht (z.B. Belmont & Butterfield 1977; Reinecker 1996; Sternberg 1983). Die Betonung wird dabei auf

”Kompetenz” gelegt, weil eine Interventionsmaßnahme gewöhnlich mehr als eine momentane Verhaltensänderung bewirken soll. Nach Hager und Hasselhorn (1997, 2000: 51) ist es sinnvoll, zwischen kurzfristig verfügbaren und langfristig verfügbaren Kompetenzen zu unterscheiden sowie den Grad an notwendigem Transfer und Generalisierbarkeit zu berücksichtigen. Sie unterscheiden zwischen dem zeitlichen Transfer, dem Situationstransfer und dem Anforderungstransfer. Mit der Reichweite des notwendigen Transfers steigen die Anforderungen an die Evaluationsforschung. Besonders bedenkenswert erscheint der zeitliche Transfer, da die meisten Interventionsmaßnahmen auf eine dauerhafte oder zumindest langfristige Wirkung angelegt sind. Darüber hinaus muss mit verzögerten Wirkungen kalkuliert werden, die nicht direkt nach Programmende beobachtet werden können. Es wird deshalb angeraten, die Wirksamkeitsmessung nicht nur auf den Zeitpunkt direkt nach Abschluss des Interventionsprogramms zu beschränken (vgl. Hager & Hasselhorn 2000: 96; Rüger & Senf 1994). Die Darstellung der Auswirkungen des Interventionsprogramms erfordert außerdem die Bestimmung des Ausgangszustands, der vor dem Einsetzen der Interventionsmaßnahme bestanden hat. Für die Kontrolle von Störfaktoren wird darüber hinaus die Einbeziehung einer Vergleichsgruppe als essentiell erachtet; dies geht aus den Vorstellungen vieler Autoren über Designs der Evaluationsforschung hervor (z.B. Baumann & Reinecker-Hecht 1998; Hager 1995; Kazdin 1980, 1994; Reinecker 1996). Insbesondere im Bereiche der Therapieevaluation hat sich die Berücksichtigung von Vergleichsgruppen aufgrund des Auftretens spontaner Remissionen als wichtig erwiesen.

In Anbetracht dieser Erfahrungen sollte nach Hager (2000: 183) ein Vortest-Nachtest-Follow-up-Vergleichsgruppen-Plan in der Evaluationsforschung den Regelfall darstellen, der sich dadurch auszeichnet, dass zu drei Zeitpunkten, nämlich vorher, nachher und mit einem zeitlichen Abstand, Daten erhoben werden sowie eine Vergleichsgruppe einbezogen wird. Weiterhin unterscheidet Hager zwischen zwei Evaluationsparadigmen: der isolierten Evaluation und der vergleichenden Evaluation. Das Evaluationsparadigma der isolierten Evaluation dient der Überprüfung der Wirksamkeitshypothese, die nur eine Interventionsmaßnahme zum Gegenstand hat; wohingegen das Evaluationsparadigma der vergleichenden Evaluation der Überprüfung der Wirksamkeitsunterschiedshypothese dient, die den Vergleich mehrerer Interventionsmaßnahmen vorsieht. In Abhängigkeit von den konkreten Anforderungen einer Evaluationsstudie kann natürlich ein Plan gewählt werden, der die Komplexität des Vortest-Nachtest-Follow-up-Vergleichsgruppen-Plans überschreitet. Er sollte jedoch nicht dahinter zurückfallen.

2.6 Die Professionalisierung der Evaluationsforschung

Es sind im wesentlichen zwei Faktoren, die eine Professionalisierung der Evaluationsforschung erforderlich werden ließen. Eine Ursache ist in der steigenden Nachfrage nach Evaluation zu sehen, die inzwischen ein beträchtliches Ausmaß erreicht hat, und eine andere im Anwachsen des Methodeninventars, dessen Umfang es wissenschaftlichen Laien in diesem Bereich zunehmend schwer macht, Evaluationsforschung zu betreiben. Obwohl Deutschland noch 1990 im Hinblick auf die Evaluationsforschung eher als ”Entwicklungsland” eingestuft wurde (vgl. Koch & Witt-

mann 1990), kann hierzulande ein zunehmendes Maß an Nachfrage beobachtet werden, auch von großen und erfolgreichen Unternehmen, wie beispielsweise Versicherungen sowie Behörden und gesetzgebenden Körperschaften. Unabhängig davon stellt das Anwachsen eines spezifischen Methodeninventars im Bereich der Evaluationsforschung zweifellos immer höhere Ansprüche an Personen, die in diesem Bereich tätig sind, so dass eine zunehmende Professionalisierung erforderlich wurde. In den Vereinigten Staaten als Ursprungsland der Evaluation war das Eintreten der Professionalisierung schon vor langer Zeit zu beobachten. In der deutschen Psychologielandschaft ist diese Professionalisierung 1987 sichtbar mit der Verabschiedung einer Rahmenprüfungsordnung für den Diplomstudiengang Psychologie eingetreten, die ein eigenes Lehr- und Prüfungsfach "Evaluation und Forschungsmethodik" vorsieht. Dieses Fach ist zwischenzeitlich an fast allen Universitäten eingeführt und wurde 1999 um ein Curriculum ergänzt (Moosbrugger, Rost & Schermelleh-Engel 1999).

3. Die Anwendungsbereiche

Eine wichtige Quelle für die Erfahrungsbildung in der Evaluation ist die Anwendungspraxis. Durch die Anwendungspraxis werden Problemstellungen vorgegeben, für die Lösungen gesucht werden müssen. Solche Erfahrungen können in die Entwicklung von Standards und Kriterien einbezogen werden, um von dort dann wieder auf die Anwendung zurückzuwirken. Darüber hinaus üben sie durch ihre Vorbildfunktion natürlich auch einen direkten Einfluss auf zukünftige Evaluationsstudien aus. Bei den Anwendungen handelt es sich allerdings nicht um einen homogenen Bereich. Von Wottawa und Thierau (1998) wird eine große Zahl spezifischer Themen aufgelistet. Wie bereits einleitend deutlich gemacht wurde, lassen sich aber mehrere übergeordnete Bereiche unterscheiden, die besonders viele Gemeinsamkeiten untereinander aufweisen (Rost 2000). Diese Bereiche werden den weiteren Überlegungen zugrunde gelegt.

3.1 Evaluation sozialer Programme

Bei der Evaluation sozialer Programme handelt es sich in der Regel um die Auseinandersetzung mit Programmen für soziale Problemgruppen wie Delinquente (z.B. Egg, Pearson, Cleland & Lipton 2001) oder Drogenabhängige (z.B. Bühringer 1997) oder auch für sozial schwache Gruppen (z.B. Frassine 1980). Eine zentrale Aufgabe der Evaluation besteht darin, den Nutzen des Programms für die Mitglieder der betroffenen Gruppen zu überprüfen. Vor allen Dingen will der Auftraggeber gewöhnlich wissen, ob die finanziellen und personellen Investitionen in das soziale Programm gerechtfertigt sind. Die Durchführung einer solchen Evaluation erfordert die Einarbeitung in den theoretischen Hintergrund des Programms, denn nach Hager, Patry und Brezing (2000) sollte die Evaluation auf der Basis der theoretisch vermuteten Wirkmechanismen geplant werden, um zu gewährleisten, dass die theoretische Konzeption und das Assessmentinstrumentarium aufeinander abgestimmt sind, damit die Chance besteht, dass ein eventueller Nutzen auch wirklich erfasst

wird. Aufgrund der für Psychologen als Evaluatoren häufig fachfremden Thematik ergeben sich dadurch besondere Anforderungen bei der Festlegung der zu erwartenden Auswirkungen und darauf abgestimmt eines geeigneten Instrumentariums für die Messung der Wirkung des Programms.

3.2 Therapieevaluation

Der Therapieevaluation kommt insbesondere im Hinblick auf die vielen verschiedenen Therapiearten, die auf einen Patienten angewendet werden können, Bedeutung zu. Diese Therapiearten unterscheiden sich hinsichtlich vieler Merkmale. Mit dieser Vielfalt kann in zwei Weisen umgegangen werden: Man kann versuchen, sie hinsichtlich ihrer durchschnittlichen Wirkung zu integrieren (Grawe 1999) oder vergleichend gegeneinander abzugrenzen. Die Therapieevaluation dient der zweiten von diesen beiden Vorgehensweisen. Sie wird eingesetzt, um diejenige(n) Therapieart(en) zu ermitteln, welche in Anbetracht der jeweiligen Patientenmerkmale den größten Nutzen erbringt (z.B. Grawe 1993). Bei der Nutzenbestimmung steht natürlich das Wohl des Patienten im Vordergrund; darüber hinaus müssen aber auch die Kosten des jeweiligen Verfahrens in Rechnung gestellt werden. In diesem Bereich von Evaluation besteht der Vorteil, dass auf originär psychologische Theorien und Messinstrumente zurückgegriffen werden kann.

3.3 Evaluation von Maßnahmen und Innovationen im Kontext der Arbeits- und Organisationspsychologie

In der Arbeits- und Organisationspsychologie werden Evaluationen bevorzugt im Kontext von Ausbildungs- und Weiterbildungsmaßnahmen eingesetzt. Sie betreffen ganz unterschiedliche Themen wie betriebliches Stressmanagementtraining (Busch 2001), polizeiliche Weiterbildung (Holling 1999) oder die Berufsbildevaluation des Datenverarbeitungskaufmanns (Hendrix 1999). Die Evaluation solcher Ausbildungs- und Weiterbildungsmaßnahmen hat häufig sowohl einen summativen als auch einen formativen Charakter. Da spezifische Erwartungen zur Wirkung der Maßnahmen bestehen, also Ziele für die Wirksamkeit vorausgesetzt werden, wird häufig die Phase der Durchführung genutzt, um die Maßnahmen im Hinblick auf die Erwartungen zu optimieren. Wie bei der Evaluation sozialer Programme ist auch hier die Thematik mitunter fachfremd und stellt deshalb hohe Anforderungen an die Festlegung der zu erwartenden Auswirkungen und an die Auswahl eines darauf abgestimmten geeigneten Instrumentariums für die Messung der Wirkung des Programms.

3.4 Evaluation im pädagogischen Bereich

Im pädagogischen Bereich steht die Evaluation von Bildungsmaßnahmen unter dem Stichwort Qualitätssicherung in Mittelpunkt. Es werden curriculare Varianten (z.B. Kersten, Groner & Stricker 2001; Seiffge-Krenke 1981), der Einsatz neuer Medien (z.B. Horz, Hofer & Fries 2001; Kokavecz, Lammers & Holling 1999), die Nutzung

von Hypermedien (z.B. Tergan 2001) und vieles mehr, was das Lernen von Schülern begünstigen soll, evaluiert. Darüber hinaus gibt es Programme, die eine Verbesserung der Randbedingungen der Ausbildung bringen sollen, wie etwa die Verringerung von Gewalt in der Schule (Atria & Spiel 2001), deren Nutzen überprüft werden muss. Insgesamt handelt es sich um einen recht homogenen Bereich, da es in jedem Fall um die Maximierung des Nutzens für die Schüler geht. Aufgrund der Nähe zu den Inhalten der Psychologie und der Verfügbarkeit einer großen Zahl von Messinstrumenten bestehen hier für die psychologische Evaluationsforschung besonders günstige Bedingungen.

4. Die Lehrevaluation

Die Lehrevaluation stellt im Hinblick auf einige Merkmale eine Besonderheit dar, die es notwendig macht, sie als eigenständigen inhaltlichen Bereich abzuhandeln. Ein erstes besonderes Merkmal besteht darin, dass bei der Evaluation insbesondere universitärer Lehre die Distanz zwischen Evaluator und Evaluiertem wesentlich geringer ist, weil Evaluatoren und Evaluierte gewöhnlich derselben Berufsgruppe angehören und nicht selten ein und dieselbe Person sind. Weitere Besonderheiten sind, dass die Lehrevaluation vergleichsweise häufig Anwendung findet und insoweit nur einen eher geringen Aufwand erfordert, als sie meist nur Einzelpersonen und deren Handeln betrifft. Aufgrund dieser Besonderheiten werden an die Forschung zur Lehrevaluation spezielle Anforderungen gestellt (Abrami & d'Apollonia 1991). So verbindet sich etwa mit der Lehrevaluation meist eine Abweichung von einem der gerade erst vorgestellten Standards, denn gewöhnlich fehlt es an einer wirklich gut geeigneten Vergleichsgruppe. Das Vorliegen einer hohen Lehrqualität muss daher auf eine andere Weise sichergestellt werden als durch den direkten Vergleich. Schließlich ist auf das besondere gesellschaftliche Interesse an der Lehrevaluation zu verweisen. Eine gute Ausbildung der nachfolgenden Generationen muss sichergestellt werden, während gleichzeitig die Kosten in Grenzen zu halten sind. Es ist daher nicht verwunderlich, dass die Lehrevaluation an privaten Lehrinstitutionen selbstverständlich ist (vgl. Conchello 1997) und die Ergebnisse in Form von Rankings der Universitäten in der Öffentlichkeit ein sehr großes Interesse gefunden haben (vgl. Dienstbühl 1997; Engel 2001; Spiel 2001).

4.1 Die Qualität der Lehre

Da der Gegenstand der Lehrevaluation die Qualität der Lehre ist, steht im Mittelpunkt der Lehrevaluation die Frage, was unter einer hohen Lehrqualität zu verstehen ist. Die Lehrqualität wird als ein Merkmal des Dozentenhandelns, das insbesondere durch didaktische Maßnahmen und sozial-interaktive Verhaltensweisen bestimmt ist, aufgefasst (Souvignier & Gold 2001). Erst wenn diese Frage in einer zufriedenstellenden Weise beantwortet werden kann, ist es möglich, Lehre theoriegeleitet zu evaluieren. Ein lediglich pragmatisches Vorgehen besteht darin, Studierende nach ihrem Urteil über die Lehre eines Dozenten zu fragen. Diese Vorgehensweise ist allerdings insofern problematisch, als dadurch Studierenden die Hauptver-

antwortung der Evaluation aufgebürdet wird, also einer Personengruppe, die selbst betroffen ist, weshalb deren Urteil in Frage gestellt werden kann (vgl. Kromrey 1994). Bei diesem Vorgehen wird den Studierenden insbesondere unterstellt, dass sie über geeignete didaktische Maßstäbe und Wissen zu den relevanten Lehrzielen wie auch zu allen für einen guten Lehrerfolg in einem Veranstaltungsverband angemessenen Bedingungen verfügen. Obwohl dieses Expertentum von Studierenden wohl fraglich ist, stellen die Studierenden eine wertvolle Informationsquelle für die Lehrevaluation dar, sofern sie in einer geeigneten Weise als Beobachter eingesetzt werden (Marsh & Roche 1997).

Die Qualität der Beobachtungen Studierender hängt, wie auch die Qualität anderer Beobachter, davon ab, ob den Beobachtungen geeignete inhaltliche Konzepte, geeignete Messkonzepte und eventuell auch eine Einführung in die Anwendung dieser Konzepte zugrunde gelegt werden (Bortz & Döring 2002). Je besser das eingesetzte Messinstrument auf die Beobachtbarkeit der relevanten Merkmale von Lehrverhalten abgestimmt ist, desto hochwertiger sind die Beobachtungen, die unter Mithilfe von Studierenden erzielt werden können. Darüber hinaus wurden von March (1987) in einer metaanalytischen Studie Merkmale von Studierenden als mögliche Ursachen für fehlerhafte Beobachtungsurteile untersucht. Dabei ergab sich, dass nur 13 Prozent der Varianz durch solche Hintergrundvariablen ("Bias-Variablen") erklärt werden konnte. Als besonders bedeutende Prädiktoren erwiesen sich das Interesse/die Erwartung an der Veranstaltung, eine hohe Erfolgserwartung, der Schwierigkeitsgrad und allgemeine Interessen. Auch die Berücksichtigung neuerer Studien führt im wesentlichen zum gleich Ergebnis (March & Roche 1997). Neben diesen auf die Studierenden bezogenen Hintergrundvariablen wurden auch auf Lehrende bezogene Hintergrundvariablen untersucht. Dabei wurde etwa festgestellt, dass Studierende besonders günstig auf eine hohe Expressivität des Lehrenden reagieren (March & Ware 1982). Aber auch dieser Effekt wird nicht als eine wesentliche Einschränkung der Güte der Beobachtungen von Studierenden betrachtet (March & Roche 1997). Auf mögliche Artefakte, die im Rahmen der studentischen Beurteilung beachtet werden müssen, haben Moosbrugger und Hartig (2001) hingewiesen.

4.2 Konzepte und Modelle für Lehrqualität

Seit dem Beginn der systematischen Auseinandersetzung mit dem Thema Lehrqualität vor ungefähr 30 Jahren haben sich vor allen Dingen die theoretischen Grundlagen und die Konzepte für die Konstruktion von Messinstrumenten geändert. Nach dem derzeitigen Stand der Forschung sollte die Modellierung von Determinanten des Studienerfolgs als Ansatzpunkt für die Entwicklung einschlägiger Konzepte gewählt werden (Souvignier & Gold 2001). Auch Theorien des Lernens sollten Berücksichtigung finden. Solche Konzepte können die bisherigen Erfahrungen mit der Lehrevaluation und didaktische Modelle zu einem integrativen Ganzen verbinden. Für den Zweck der Lehrevaluation wird von Souvignier und Gold eine Bezugnahme auf die Modelle von Helmke (1996) und Rindermann (1999) sowie auf das Lernprozessmodell von Steiner (1999) empfohlen. Das "Angebots-Nutzungs-Modell" von Helmke betont die Notwendigkeit des Zusammenpassens von Merk-

malen der Lehrveranstaltung auf der einen Seite und den Voraussetzungen der Studierenden auf der anderen Seite. Rindermanns "Münchner multifaktorelles Modell der Lehrveranstaltungsqualität" unterscheidet mehrere Wirkfaktoren und betont ebenfalls den Aspekt der Passung zwischen Dozentenangebot und studentischen Voraussetzungen. Dagegen stehen bei Steiners Lernprozessmodell affektiv-motivationale Überlegungen im Vordergrund, die den Aufbau und Erhalt von Lernbereitschaft und Akzeptanz der Lernziele fördern sollen.

4.3 Messinstrumente für Lehrqualität

Einen sehr differenzierten Überblick über Messinstrumente zur Vorlesungsevaluation bieten Souvignier und Gold (2001). Sie machen deutlich, dass eine Auswahl unterschiedlicher Fragebogen mit einer guten Qualität für diesen Zweck verfügbar ist, und unterscheiden zwischen drei Typen von Fragebogen für Vorlesungen: (1) kurze Fragebögen mit dem Anspruch einer wenig differenzierten Erfassung der Lehrqualität, (2) Fragebögen für eine differenzierte Erfassung des Dozentenverhaltens und der Aufbereitung des Lehrstoffs, (3) Fragebögen für eine umfassende Beurteilung der Qualität universitärer Lehre. Der letztere Typ von Fragebogen beansprucht, die Lehre in allen relevanten Aspekten zu erfassen. Die Testkennwerte der Fragebögen, die nach der Klassischen Testtheorie konstruiert wurden, weisen auf eine hohe Testgüte hin.

Mit der Suche nach einem angemessenen Konzept stellt sich allerdings auch die Frage nach einem geeigneten Differenzierungsgrad für die Repräsentation von Lehrqualität. In dieser Frage bestehen derzeit noch recht unterschiedliche Auffassungen. Der Verschiedenheit der Fragebogen, die zur Lehrevaluation eingesetzt werden, ist zu entnehmen, dass es Vertreter der Position einer einzigen oder doch geringen Anzahl von Dimensionen gibt; ihnen stehen Vertreter einer vieldimensionalen Auffassung gegenüber (vgl. Abrami, d'Apollonia & Rosenfield 1997).

Für die Erfassung der Lehrqualität wurde eine große Zahl von Messinstrumenten entwickelt, die auf teilweise unterschiedlichen theoretischen Positionen beruhen. Bezüglich der Darbietungsform kann zwischen den traditionellen Fragebogen (vgl. Souvignier & Gold 2001), computergestützten Methoden (vgl. Meyer 1997) und interaktiven Methoden (vgl. Hartig 1997) unterschieden werden. Weiterhin wurden Messinstrumente für bestimmte Veranstaltungstypen entwickelt. Für die Evaluation von Vorlesungen liegen spezielle Fragebogen vor (z.B. Diehl & Kohr 1977; Rindermann 2001; Westermann, Spies, Heise & Wollburg-Claar 1998) ebenso wie für Seminare (z.B. Gold & Souvignier 2000; Staufenbiel 2000). Insgesamt erfassen diese Fragebogen das Dozentenverhalten anhand von Struktur und Organisation der Lehrveranstaltung sowie in Bezug auf die Motivierungsqualität und sozial-interaktive Anteile. Auch die Kompetenz des Dozenten wird angesprochen, das Anforderungsniveau und die Auswahl der Lehrinhalte. In den Fragebögen von Moosbrugger, Naumann und Hartig (1997) und Moosbrugger, Hartig und Struwe (1999) wird zusätzlich auch das studentische Lernverhalten explizit erfasst. Die verschiedenen Erhebungsinstrumentarien berücksichtigen je nach Konzeption viele oder wenige dieser Aspekte mit einem großen oder geringen Differenzierungsgrad. Manche Fragebogen liegen bereits in mehreren Versionen vor (z.B. Rindermann 1996, 2001).

Die Befragung der Studierenden ist zwar sehr wesentlich; sie stellt aber nicht die einzige Informationsquelle für die Lehrevaluation dar. Andere wichtige Informationen, die zur Beurteilung der Lehre herangezogen werden können, sind die Zahl der Studierenden, die ein bestimmtes Lehrangebot wahrnehmen, der Schwund an Veranstaltungsteilnehmern, der im Verlaufe eines Semesters beobachtet werden kann, die Anzahl der betreuten Hausarbeiten, Abschlussarbeiten und anderer Produkte. Außerdem bietet es sich an, bei der Lehrevaluation die jeweiligen Randbedingungen in Rechnung zu stellen. So kann sich ein ungünstiger Veranstaltungszeitpunkt oder ein ungünstiger, schlecht ausgestatteter Veranstaltungsraum durchaus negativ auf die Bewertung auswirken. Solche Bedingungen, die auf die Beurteilung der Lehrqualität einen (wenn auch nur geringen) Einfluss nehmen können, sind neben Fragen zur Motivation und dem verfügbaren Zeitbudget etc. mit dem Studien-Bedingungs-Fragebogen (Moosbrugger, Struwe, Hartig & Reiß 1999) für die Lehrevaluation erfassbar.

4.4 Konsequenzen der Messung von Lehrqualität

Abschließend ist noch darauf hinzuweisen, dass die Lehrevaluation sich nicht nur auf die Erhebung des Ist-Zustandes beschränken sondern auch für Konsequenzen genutzt werden sollte (Rindermann 2001). Insbesondere sollten die Evaluierten aus dem Ergebnis der Lehrevaluation Nutzen im Sinne einer Verbesserung ihrer Lehre ziehen. Es wird davon ausgegangen, dass die unvermittelte Weitergabe der Ergebnisse diesen Zweck nicht erfüllen kann. Statt dessen wird empfohlen, die Vermittlung der Ergebnisse um Erläuterungen oder gar eine gezielte Beratung zu ergänzen (Henning & Balk 2001; Marsh 1987; Marsh & Roche 1997; Rindermann 2001). Weiterhin besteht die Möglichkeit, durch Schulungsmaßnahmen die Kompetenz der Dozenten gegebenenfalls in geeigneter Weise zu verbessern (z.B. Berendt 2000; Webler 2000; Winteler & Krapp 1999). Neben Rückmeldung und eventueller Modifikation als wichtigsten Zielsetzungen der Lehrevaluation werden die Ergebnisse natürlich auch noch für die Forschung und für administrative Zwecke verwendet.

5. Schluss

In der Psychologie wird das Thema Evaluation aus vielen Perspektiven bearbeitet, insbesondere aus der Perspektive der Methodik. Es ist eine dynamische Entwicklung zu beobachten, die ihren Nährboden in der Interaktion zwischen den Evaluationsmethoden und den Anwendungsgebieten findet. Insgesamt besteht die Tendenz, die Qualität der Evaluation im Zuge der psychologischen Evaluationsforschung fortlaufend zu verbessern.

Literatur

Abrami, P. C. & d'Apollonia: (1991): Multidimensional students' evaluation of teaching effectiveness – generalizability of "N=1" research: Comments on March (1991). In: Journal of Educational Psychology, 30: 221-227.

Abrami, P. C., d'Apollonia, S. & Rosenfield: (1997): The dimensionality of student ratings of instruction. What we know and what we do not. In: R. P. Perry & J. C. Smart (Hg.): Effective teaching in higher education: Research and practice. New York: Agathon Press: 321-367.

Atria, M. & Spiel, C. (2001): Programmevaluation im Bildungsbereich – von der Schwierigkeit, Effekte zu messen. In: H. Moosbrugger, K. Schermelleh-Engel, J. Hartig & Y. Brandl (Hg.): Methoden & Evaluation. Tagungsband der 5. Tagung der Fachgruppe Methoden und Evaluation der Deutschen Gesellschaft für Psychologie. Frankfurt a. M.: Fachbuchhandlung für Psychologie.

Baumann, U. & Reinecker-Hecht, C. (1998): Methodik der klinisch-psychologischen Interventionsforschung. In U. Baumann & M. Perrez (Hg.): Lehrbuch Klinische Psychologie – Psychotherapie (2. Aufl.: 346-365). Bern: Huber.

Berendt, B. (2000): Was ist gute Hochschullehre? In: Zeitschrift für Pädagogik, 41. Beiheft: 247–259.

Belmont, J. M. & Butterfield, E. C. (1977): The instructional approach to developmental cognitive research. In: R. V. Kail & J. W. Hagen (Hg.): Perspectives on the development of memory and cognition. Hilsdale, NJ: Erlbaum: 437-481.

Bortz, J. & Döring, N. (2002): Forschungsmethoden und Evaluation (3. Aufl.). Berlin: Springer.

Bredenkamp, J. (1980): Theorie und Planung psychologischer Experimente. Darmstadt: Steinkopff.

Bühringer, G. (1997): Versorgungsstruktur und Qualitätssicherung in der Suchthilfe. In: J. Ulrich (Hg.), Deutsche Hauptstelle gegen die Suchtgefahren Regionale Suchtkrankenversorgung. Freiburg: Lambertus.

Bunge, M. (1967): The search for truth (Scientific research, Vol. II). Berlin: Springer.

Campbell, D. T. & Stanley, J. C. (1963): Experimental and quasi-experimental designs for research and teaching. In: N. L. Gage (Ed.): Handbook for research and teaching. Chicago: Rand McNally.

Campbell, D. T. & Stanley, J. C. (1973): Experimentelle und quasi-experimentelle Anordnungen in der Unterrichtsforschung. In: K. Ingenkamp (Hg.): Strategien der Unterrichtsforschung (Teilausgabe des Handbuches der Unterrichtsforschung, S.99-193). Beltz: Weinheim.

Conchello, M. (1997): Lehrevaluation an privaten Hochschulen am Beispiel der Europäischen Wirtschaftshochschule Paris Oxford Berlin Madrid (E.A.P). In: H. Moosbrugger & D. Frank (Hg.): Möglichkeiten und Grenzen der wissenschaftlichen Evaluation universitärer Lehre. Riezlern-Reader V (S. 141-148). Arbeiten aus dem Institut für Psychologie der Johann Wolfgang Goethe-Universität, Frankfurt a. M.

Cook, T. D. & Campbell, D. T. (1979): Quasi-experimentation. Design and analysis of issues for field setting. Boston, MA: Houghton Mifflin.

Cook, T. D., Campbell, D. T. & Perachio, L. (1990): Quasi-experimentation. In: M. D. Dunnette & L. M. Hough (Hg.): Handbook of industrial and organizational psychology (2nd ed., Vol. 1, S.491-576). Palo Alto, CA: Consulting Psychologists Press.

Cook, T. D. & Matt, G. E. (1990): Theorien der Programmevaluation – Ein kurzer Abriß. In: U. Koch & W. W. Wittmann (Hg.): Evaluationsforschung. Bewertungsgrundlage von Sozial- und Gesundheitsprogrammen. Berlin: Springer.

Diehl, J.M. & Kohr, H.-U. (1977): Entwicklung eines Fragebogens zur Beurteilung von Hochschulveranstaltungen im Fach Psychologie. Psychologie in Erziehung und Unterricht, 24: 61-75.

Deutsche Gesellschaft für Evaluation e. V. (2002): Standards für Evaluation. Köln.

Dienstbühl, I. (1997): Inhaltliche und methodische Grundlagen des „Spiegel"-Universitätsrankings. In: H. Moosbrugger & D. Frank (Hg.): Möglichkeiten und Grenzen der wissenschaftlichen Evaluation universitärer Lehre. Riezlern-Reader V (S. 149-170). Arbeiten aus dem Institut für Psychologie der Johann Wolfgang Goethe-Universität, Frankfurt a. M.

Ebbinghaus, H. (1908): Abriß der Psychologie. Leipzig: Veit.
Egg, R., Pearson, F. S., Cleland, M. & Lipton, D. S. (2001): Evaluation von Straftäterbehandlungsprogrammen in Deutschland: Überblick und Meta-Analyse. Pfaffenweiler: Centaurus.
Engel, U. (2001): Hochschulranking. Zur Qualitätsbeurteilung von Studium und Lehre. Frankfurt: Campus Verlag.
Frassine, J. (1980): Evaluation von sozialen Modellen auf Klientenebene. In: Österreichische Zeitschrift für Soziologie, 5: 53-58.
Gold, A. & Souvignier, E. (2000): Rückmeldegespräche nach studentischen Referaten. Ein Beitrag zur Verbesserung von Lehre? In: G. Krampen & H. Zayer (Hg.): Psychologiedidaktik und Evaluation II. Bonn: Deutscher Psychologen Verlag: 203-218.
Grawe, K. (1993): Über Voraussetzungen eines gemeinsamen Erkenntnisprozesses in der Psychotherapie. In: Psychologische Rundschau, 44: 181-186.
Grawe, K. (1999): Gründe und Vorschläge für eine Allgemeine Psychotherapie. Psychotherapeut, 44: 350-359.
Hager, W. (1987): Grundlagen der Versuchsplanung zur Prüfung empirischer Hypothesen in der Psychologie. In: G. Lüer (Hg.): Allgemeine experimentelle Psychologie. Stuttgart: Fischer: 43-264.
Hager, W. (1995): Planung und Durchführung der Evaluation von kognitiven Förderprogrammen. In: W. Hager (Hg.): Programme zur Förderung des Denkens bei Kindern. Konstruktion, Evaluation und Metaevaluation. Göttingen: Hogrefe: 100-206.
Hager, W. (1998): Zur Validität pädagogisch-psychologischer Versuche. Empirische Pädagogik, 12: 167-210.
Hager, W. (2000): Wirksamkeits- und Wirksamkeitsunterschiedshypothesen, Evaluationsparadigmen, Vergleichsgruppen und Kontrolle. In: W. Hager, J.-L. Patry & H. Brezing (Hg.): Evaluation psychologischer Interventionsmaßnahmen. Bern: Huber: 180-201.
Hager, W. & Hasselhorn, M. (2000): Psychologische Interventionsmaßnahmen: Was sollen sie bewirken können? In: W. Hager, J.-L. Patry & H. Brezing (Hg.): Evaluation psychologischer Interventionsmaßnahmen. Bern: Huber: 41-85.
Hager, W., Patry, J.-L. & Brezing, H. (2000): Einleitung und Überblick. In: W. Hager, J.-L. Patry & H. Brezing (Hg.): Evaluation psychologischer Interventionsmaßnahmen. Bern: Huber: 1-7.
Hager, W. & Westermann, R. (1983): Planung und Auswertung von Experimenten. In: J. Bredenkamp & H. Feger (Hg.): Hypothesenprüfung (Enzyklopädie der Psychologie, Themenbereich Methodologie und Methoden, Serie Forschungsmethoden der Psychologie, Bd. 5: 24-238). Göttingen: Hogrefe.
Hartig, J. (1997): Dialogorientierte Evaluation als Alternative zu quantitativen Erhebungsverfahren. In: H. Moosbrugger & D. Frank (Hg.): Möglichkeiten und Grenzen der wissenschaftlichen Evaluation universitärer Lehre. Riezlern-Reader V. Arbeiten aus dem Institut für Psychologie der Johann Wolfgang Goethe-Universität, Frankfurt a. M: 112-119.
Helmke, A. (1996): Studentische Evaluation der Lehre – Sackgassen und Perspektiven. In: Zeitschrift für Pädagogische Psychologie, 10: 181-186.
Hendrix, W. (1999): Evaluation des Berufsbildes des Datenverarbeitungskaufmanns. In: H. Holling & Günther Gediga (Hg.). Evaluationsforschung. Göttingen: Hogrefe: 127- 178.
Henninger, M. & Balk, M. (2001): Return to sender – die Rückmeldung von Schulleistungsdaten als Forschungsgegenstand. In: H. Moosbrugger, K. Schermelleh-Engel, J. Hartig & Y. Brandl (Hg.): Methoden & Evaluation. Tagungsband der 5. Tagung der Fachgruppe Methoden und Evaluation der Deutschen Gesellschaft für Psychologie. Frankfurt a. M.: Fachbuchhandlung für Psychologie.
Henerson, M. E., Morris, L. L. & Fitz-Gibbon, C. T. (1987): How to measure attitudes. Newbury Park: Sage Publications.
Holling, H. (1999): Evaluation eines komplexen Fortbildungsprogramms zur Steigerung der beruflichen Kompetenz. In: H. Holling & G. Gediga (Hg.): Evaluationsforschung. Göttingen: Hogrefe: 1- 33.

Horz, H., Hofer, M. & Fries: (2001): Multimethodale Evaluation virtueller Hochschullehre. In: H. Moosbrugger, K. Schermelleh-Engel, J. Hartig & Y. Brandl (Hg.): Methoden & Evaluation. Tagungsband der 5. Tagung der Fachgruppe Methoden und Evaluation der Deutschen Gesellschaft für Psychologie. Frankfurt a. M.: Fachbuchhandlung für Psychologie.

Joint Commitee on Standards for Educational Evaluation (1994): The program evaluation standards. How to assess evaluations of educational programs (2nd ed.). Thousand Oaks, CA: Sage.

Kazdin, A. E. (1980): Research design in clinical psychology. New York: Harper & Row.

Kazdin, A. E. (1994): Comparative outcome studies of psychtherapy: Methodological issues and strategies. In: Journal of Consulting and Clinical Psychology, 54: 95-105.

Kersten, B., Groner, R., Groner, M. & Stricker, D. (2001): Evaluation des Projekts „Methodological Education for the Social Sciences“ im Rahmen des virtuellen Campus Schweiz. In: H. Moosbrugger, K. Schermelleh-Engel, J. Hartig & Y. Brandl (Hg.): Methoden & Evaluation. Tagungsband der 5. Tagung der Fachgruppe Methoden und Evaluation der Deutschen Gesellschaft für Psychologie. Frankfurt a. M.: Fachbuchhandlung für Psychologie.

Koch, U. & Wittmann, W. W. (1990): Einige Forderungen für die Weiterentwicklung der Evaluationsforschung in der Bundesrepublick Deutschland. In: U. Koch & W. W. Wittmann (Hg.): Evaluationsforschung. Bewertungsgrundlage von Sozial- und Gesundheitsprogrammen. Berlin: Springer.

Kokavecz, I., Lammers, F. & Holling, H. (1999): Evaluation von computergestützten Lern- und Lehrprojekten. In: H. Holling & Günther Gediga (Hg.): Evaluationsforschung. Göttingen: Hogrefe: 59- 72.

Kromrey, H. (1994): Evaluation der Lehre durch Umfrageforschung? Methodische Fallstricke bei der Messung von Lehrqualität durch Befragung von Vorlesungsteilnehmern. In: P. Ph. Mohler (Hg.): Universität und Lehre. Ihre Evaluation als Herausforderung an die empirische Sozialforschung. Münster: Waxmann.

Marsh, H. W. (1987): Student’s evaluations of university teaching: Research findings, methodological issues, and directions for future research. In: International Journal of Educational Research, 11 (Whole Issue No. 3).

Marsh, H. W. & Roche, L. (1997): Making students' evaluations of teaching effectiveness effective. In: American Psychologist, 52: 1187-1197.

March, H. W. & Ware, J. E. (1982): Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. In: Journal of Educational Psychology, 74: 126-134.

Meyer, A. (1997): PC-gestützte Fragebogensysteme zur Hochschulevaluation am Beispiel der Fachhochschule Heilbronn. In: H. Moosbrugger & D. Frank (Hg.): Möglichkeiten und Grenzen der wissenschaftlichen Evaluation universitärer Lehre. Riezlern-Reader V. Arbeiten aus dem Institut für Psychologie der Johann Wolfgang Goethe-Universität, Frankfurt a. M: 120-140.

Mittag, W. & Hager, W. (2000): Ein Rahmenkonzept zur Evaluation psychologischer Interventionsmaßnahmen. In: W. Hager, J.-L. Patry & H. Brezing (Hg.): Evaluation psychologischer Interventionsmaßnahmen. Bern: Huber: 102-128.

Mittag, W. & Jerusalem, M. (1997): Evaluation von Präventionsprogrammen. In: R. Schwarzer (Hg.): Gesundheitspsychologie. Ein Lehrbuch (2. Aufl., S.595-611). Göttingen: Hogrefe.

Moosbrugger, H. & Hartig, J. (2001): Zur Bedeutung von individuellen und institutionellen Studienbedingungen für die vergleichende Evaluation der Lehre. In: U. Engel (Hg.): Hochschulranking. Zur Qualitätsbeurteilung von Studium und Lehre. Frankfurt: Campus Verlag.

Moosbrugger, H., Hartig, J. & Struwe: (1999): Fragebogen zum Lehr- und Lernverhalten – Allgemein (FELL-A). (Arbeiten aus dem Institut für Psychologie der Johann Wolfgang Goethe-Universität Heft 11). Frankfurt a. M.: Universität, Institut für Psychologie.

Moosbrugger, H., Naumann. J. & Hartig, J. (1997): Fragebogen zur Evaluation des lehr- und Lernverhaltens (FELL). (Arbeiten aus dem Institut für Psychologie der Johann Wolfgang Goethe-Universität Heft 5). Frankfurt a. M.: Universität, Institut für Psychologie.

Moosbrugger, H., Rost, J. & Schermelleh-Engel, K. (1999): Überlegungen zur Entwicklung eines Curriculums für das Prüfungsfach ”Evaluation und Forschungsmethoden”. In: Psychologische Rundschau, 50: 165-167.

Moosbrugger, M., Struwe, S., Hartig, J. & Reiß: (1999): Studien Bedingungs- Fragebogen. (Arbeiten aus dem Institut für Psychologie, Heft 12/1999). Frankfurt am Main: J. W. Goethe-Universität, Institut für Psychologie.

Morris, L. L., Fitz-Gibbon, C. T. & Lindheim, E. (1987): How to measure performance and use tests. Newbury Park: Sage Publications.

Patry, J.-L. & Perrez, M. (2000): Theorie-Praxis-Probleme und die Evaluation von Interventionsprogrammen. In: W. Hager, J.-L. Patry & H. Brezing (Hg.): Evaluation psychologischer Interventionsmaßnahmen. Bern: Huber: 19-40.

Reinecker, H. (1996): Therapieforschung. In: J. Margraf (Hg.): Lehrbuch der Verhaltenstherapie (Bd. 1: Grundlagen – Diagnostik – Verfahren – Rahmenbedingungen: 31-48). Berlin: Springer.

Rindermann, H. (1996): Untersuchungen zur Brauchbarkeit studentischer Lehrevaluationen. Landau: Verlag Empirische Pädagogik.

Rindermann, H. (1999): Bedingungs- und Effektvariablen in der Lehrevaluationsforschung: Konzeption und Prüfung des Münchner multifaktoriellen Modells der Lehrveranstaltungsqualität. Unterrichtswissenschaft, 27: 357-380.

Rindermann, H. (2001): Lehrevaluation. Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen. Landau: Verlag für Empirische Pädagogik.

Rossi, P. H. (1982): Standards for evaluation practice. San Francisco: Jossey-Bass.

Rossi, P. H. (1984): Professionalisierung der Evaluationsforschung? Beobachtungen zu Entwicklungstrends in den USA. In: G.-M. Hellstern & H. Wollmann (Hg.): Handbuch zur Evaluierungsforschung (Bd. 1). Opladen: Westdeutscher Verlag GmbH.

Rossi, P. H. & Freeman, H. E. (1993): Evaluation. A systematic approach (5th ed.). Newburry Park, CA: Sage.

Rossi, P. h., Freeman H. E. & Hofmann, G. (1988): Programm-Evaluation: Einführung in die Methoden angewandter Sozialforschung. Stuttgart: Enke.

Rost, J. (2000): Allgemeine Standards für Evaluationsforschung. In: W. Hager, J.-L. Patry & H. Brezing (Hg.): Evaluation psychologischer Interventionsmaßnahmen. Bern: Huber: 129-140.

Rüger, U. & Senf, W. (1994): Evaluative Psychotherapieforschung: Klinische Bedeutung von Psychotherapie-Katamnesen. In: Zeitschrift für Psychosomatische Medizin und Psychoanalyse, 40: 103-116.

Sarris, V. (1990): Methodische Grundlagen der Experimentalpsychologie, Bd 1 und 2. München: Reinhardt.

Schiffler, A. & Hübner: (2000): Allgemeine Standards für die Evaluationspraxis. Die Standards des "Joint Commitee on Standards for Educational Evaluation" und ihre Anwendung auf praktische Aspekte bei der Evaluation von psychologischen Interventionsmaßnahmen. In: W. Hager, J.-L. Patry & H. Brezing (Hg.): Evaluation psychologischer Interventionsmaßnahmen. Bern: Huber: 141-153.

Scriven, M. (1967): The methodology of evaluation. In: R. W. Tyler, R. M. Gagné & M. Scriven (Hg.): Perspectives of curriculum evaluation. Chicago : Rand McNally: 39-83.

Scriven, M. (1972): The Methodologie der Evaluation. In: C. Wulf (Hg.): Evaluation. Beschreibung und Bewertung von Unterricht, Curricula und Schulversuchen. München : Piper: 60-91.

Scriven, M. (1991): Evaluation thesaurus (4th ed.). Newburry Park, CA : Sage.

Seiffge-Krenke, I. (1981): Handbuch Psychologieunterricht. Bd. 1 & 2. Düsseldorf: Pädagogischer Verlag Schwann.

Souvignier, E. & Gold, A. (2001): Instrumente zur Lehrevaluation und deren Potenzial zur Verbesserung der universitätren Lehre. Manuskript zur Publikation eingereicht.

Spiel, C. (2001): Evaluation universitärer Lehre – zwischen Qualitätsmanagement und Selbstzweck. Münster: Waxmann.

Staufenbiel, T. (2000): Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende. In: Diagnostica, 46: 169-181.

Steiner, G. (1999): Aspekte der Hochschuldidaktik. Info ZHW (Züricher Hochscule Winterthur), In: Sonderheft "Unterricht und lehre", 6/99: 12-21.

Stecher, B. M. & Davis, W. A. (1987): How to focus an evaluation. Newbury Park: Sage Publications.

Sternberg, R. J. (1983): Criteria for intellectual skills training. In: Educational Researcher, 13: 6-12, 26.

Tergan: O. (2001): Lernen und Wissensmanagement mit Hypermedien: Zwei Perspektiven der Qualitätsbeurteilung. In: H. Moosbrugger, K. Schermelleh-Engel, J. Hartig & Y. Brandl (Hg.): Methoden & Evaluation. Tagungsband der 5. Tagung der Fachgruppe Methoden und Evaluation der Deutschen Gesellschaft für Psychologie. Frankfurt a. M.: Fachbuchhandlung für Psychologie.

Webler, W.-D. (2000): Weiterbildung der Hochschullehrer als Mittel der Qualitätssicherung. In: Zeitschrift für Pädagogik, 41, Beiheft: 225-246.

Westermann, R. (1987): Strukturalistische Therapiekonzeption und empirische Forschung in der Psychologie. Berlin: Springer.

Westermann, R. (2002): Wissenschaftstheorie und Experimentalmethodik. Ein Lehrbuch zur psychologischen Methodenlehre. Göttingen: Hogrefe.

Westermann, R., Spies, K., Heise, E. & Wollburg-Claar: (1998): Bewertung von Lehrveranstaltungen und Studienbedingungen durch Studierende: Theorieorientierte Entwicklung von Fragebögen. In: Empirische Pädagogik, 12: 133-166.

Winteler, A. & Krapp, A. (1999): Programme zur Förderung der Qualität der Lehre an Hochschulen. In: Zeitschrift für Pädagogik, 45: 45-60.

Wittmann, W. W. (1985): Evaluationsforschung. Heidelberg: Springer.

Wottawa, H. & Thierau, M. (1998): Lehrbuch Evaluation (2. Aufl.). Göttingen: Hogrefe.