7. PROBLEME DER EVALUATIONSFORSCHUNG

Unter Evaluation versteht man allgemein die Bewertung eines Produktes, Prozesses oder eines Programms (vgl. Wittmann, 1990). Evaluationsforschung bezeichnet im Gegensatz zum Begriff Evaluation nur solche Bewertungsprozesse, in denen systematisch wissenschaftliche Forschungsmethoden eingesetzt werden. Rossi, Freeman und Hofmann (1988, S. 3) definieren als Evaluationsforschung „eine systematische Anwendung sozialwissenschaftlicher Forschungsmethoden zur Bewertung der Konzeption, Ausgestaltung, Umsetzung und des Nutzens“ sozialer Interventionsprogramme.

Evaluation entwickelte sich im Zusammenhang mit der Ausbreitung sozialwissenschaftlicher Methoden, Forschungen und Interventionen im zwanzigsten Jahrhundert. Wichtigster Stimulus in den 60er Jahren war die Initiierung sozialer Interventions- und Reformprogramme unter Kennedy und Johnson, die zu einer starken Expansion von Evaluationsstudien führten. Cook und Matt (1990) skizzieren ab diesem Zeitpunkt drei Epochen in der Geschichte der Programmevaluation:

Die erste Epoche von 1965 bis 1975 charakterisieren sie als „Dominanz objektivistischer Evaluationsmodelle“. Das Vorgehen war streng experimentell-wissenschaftlich, der Schwerpunkt lag auf der Entdeckung von Programmeffekten. Bedeutende Namen dieser Epoche waren Michael Scriven und Donald Campbell.

Die zweite Periode dauerte etwa von 1975 bis 1982 und ist als eine Gegenbewegung zum Primat quantitativer Forschung zu verstehen. Qualitative Forschung wurde gegenüber quantitativer Forschung als wichtiger eingestuft, Prozesse wurden betont und die Informationsbedürfnisse der beteiligten Gruppen als vorrangig erachtet.

Die dritte Epoche und letzte Periode schließlich nach 1980 umfasst Synthese- und Reformversuche. Die bekanntesten Vertreter dieser Epoche sind Rossi und Cronbach. Mit den Arbeiten Rossi’s (Rossi & Freeman, 1982; Chen & Rossi, 1980, 1983) sind die beiden Konzepte „multi-goal“ und „theory driven“ eng verbunden. Rossi forderte, dass mit einer Evaluation - abhängig von den Interessen der Beteiligtengruppen - gleichzeitig verschiedene Fragestellungen beantwortet werden und damit verschiedene Zielsetzungen erfüllt werden sollten („multi-goal“). Die zweite Aufgabe sieht Rossi in der Entwicklung von Theorien sozialer Interventionsprogramme. Die theoretische Fundierung der Programme soll einen wesentlichen Stellenwert für die Begründung der Programme spielen. Daher sieht Rossi die Aufgabe des Evaluationsforschers auch darin, diese latente Theorie zu identifizieren und innerhalb der Evaluation einer Überprüfung zu unterziehen („theory driven“). Die vorliegende Evaluationsstudie fühlt sich am ehesten dem Ansatz Rossi’s verpflichtet. Kritisch anzumerken zu Rossi’s Zielsetzung der Theorieüberprüfung ist, dass Grundlagenwissen nicht eins zu eins in Anwender- oder Praxiswissen übertragen werden kann (vgl. Brandtstädter, 1990, von Kardoff, 1993), sondern allenfalls Aussagen bezüglich der Effizienz dieser Theorie für bestimmte Anwendungsbereiche getroffen werden können. Hager (1998) ist gar der Meinung, dass die den Programmen zugrundeliegenden Theorien nicht mitgeprüft werden können. Er sieht den Zweck von Theorien bei pädagogischen Interventionsprogrammen darin, dass diese helfen, Programmwirkungen zu erklären.

7.1. Arten der Evaluationsforschung

Evaluationsforschung kann sich auf verschiedene Stadien eines Programms beziehen und somit verschiedene Ziele verfolgen. Bereits Scriven (1967) unterschied die formative von der summativen Evaluation. Während es bei der formativen Evaluation um eine Bewertung der Programmkonzeption und der Durchführung des Programms geht, steht bei der summativen Evaluation die Feststellung der Programmwirkung im Mittelpunkt. Bisher dominiert im Bereich der sozialen Interventionsprogramme, aber auch der Gesundheitsprogramme die Ergebnisevaluation. Dies legen zumindest die in diesem Bereich publizierten Ergebnisse nahe (vgl. auch Dlugosch und Wottawa, 1994). Möglicherweise spielt hier auch der Wunsch der Auftraggeber nach einer möglichst eindeutigen Informationsbasis für die anstehenden Entscheidungen eine Rolle. Es wird in letzter Zeit jedoch zunehmend kritisiert, dass bei der Bewertung von Programmen Bedingungen, die zum Entstehen der Programmwirkungen beigetragen haben, zu wenig berücksichtigt werden (vgl. Mittag & Jerusalem, 1997; von Kardoff, 1993). Auf die spezifischen Probleme, die dadurch entstehen, soll nachfolgend noch etwas genauer eingegangen werden.

Auch Rossi, Freeman und Hofmann (1988) unterscheiden in Abhängigkeit von Stadien des Evaluationsprojektes drei Hauptarten der Evaluationsforschung:

1) Evaluation der Programmkonzeption

In diesem Stadium geht es zunächst darum, die Konzeption von geplanten Maßnahmen bereits vor deren Durchführung zu bewerten. Hier sollte z.B. überlegt werden, ob eine Übereinstimmung zwischen Programmzielen und Programmaßnahmen besteht, inwieweit die geplanten Maßnahmen angemessen und umsetzbar sind und in welchem Verhältnis der erwartete Nutzen zu den erwarteten Kosten steht.

2) Evaluation der Programmdurchführung

Diese Evaluationsart wird auch als Prozessevaluation bezeichnet. Hierbei geht es um die kontinuierliche Überwachung der Umsetzung und Ausführung der geplanten Maßnahmen. Zum einen können so potentielle Bedingungen für später fehlende Programmwirkungen aufgedeckt werden, wenn zum Beispiel Interventionsmaßnahmen nicht regelgeleitet durchgeführt werden. Darüber hinaus besteht so die Möglichkeit frühzeitig unerwünschte Nebenwirkungen festzustellen, die die Konzeption und Wirkung eines Programms in Frage stellen oder diese vermindern können.

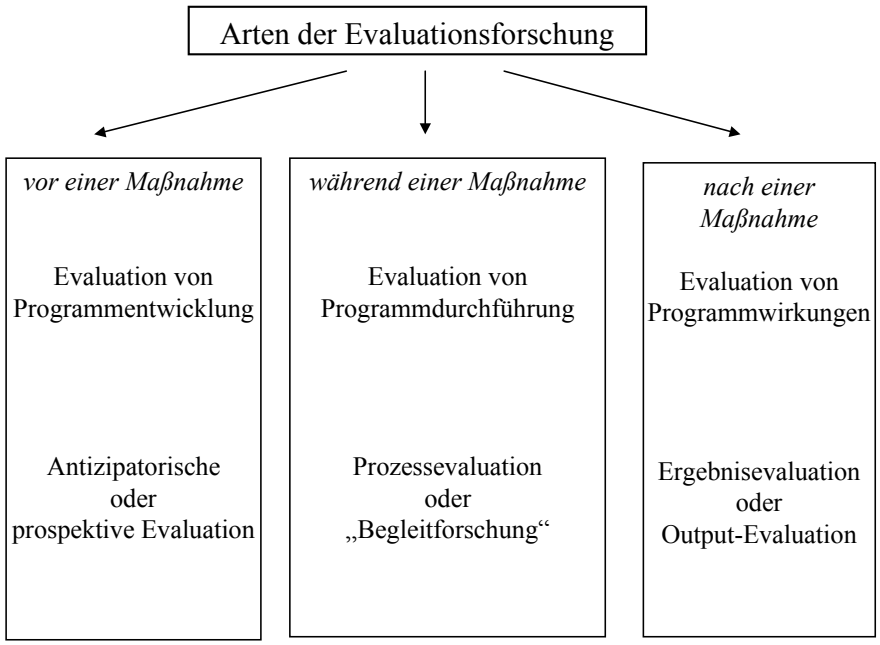

Abbildung 7-1. Arten der Evaluationsforschung

3) Evaluation der Programmwirkungen

Hierbei geht es um die Ermittlung der Wirksamkeit eines Programms, aber auch des Verhältnisses zwischen Kosten und Nutzen (Effizienz). Bei der Ermittlung der Wirksamkeit des Programms ist auch die Frage der Kausalität angesprochen, d.h. die Frage, ob die Effekte auch durch andere, als in dem Programm untersuchten Ursachenkomplexe erklärt werden könnten. Da die soziale Realität sehr komplex ist, kommen prinzipiell neben den Programmaßnahmen immer auch weitere Faktoren und Störvariablen als Ursache für Programmwirkungen in Frage. Gerade bei Präventionsprogrammen, bei denen in der Regel eher geringe Wirkungen und Veränderungen zu erwarten sind und die zudem möglicherweise nur sehr schwer nachzuweisen sind (Mittag & Jerusalem, 1997), sind systematische Planungen von Nöten, die Fehlerquellen kontrollieren und alternative Erklärungsmöglichkeiten ausschließen helfen (vgl. die Systematik der Validitätsbedrohung Campbell & Stanley [1963] sowie Cook und Campbell, [1979]).

Idealerweise sollte ein Evaluationsvorhaben alle drei Evaluationsarten umfassen. In der Realität werden abhängig von finanziellen Mitteln und Zielsetzung häufig Schwerpunkte im einen oder anderen Bereich gesetzt. Wie schon erwähnt wurden Gesundheitsprogramme am häufigsten im Hinblick auf ihre Programmwirkungen evaluiert. Mittag und Jerusalem (1995, 1997) empfehlen jedoch in Anlehnung an Rossi und Freeman (1993) Maßnahmen der Programmentwicklung, Programmdurchführung und Wirkungskontrolle aufeinander abzustimmen und als verzahnte Aktivitäten eines umfassenden Evaluationsvorhabens durchzuführen („comprehensive evaluation“).

7.2. Interne und externe Validität

Aus der Sichtweise der positivistischen quantitativ-experimentellen Forschungstradition sind experimentelle oder quasi-experimentelle Versuchspläne unabdingbare Voraussetzung, um zu methodisch abgesicherten Ergebnissen zu gelangen. So wird zur Kontrolle von Störvariablen ein Untersuchungsdesign mit mindestens einer Interventions- und einer Kontrollgruppe, einem Prä- und Posttest, sowie möglichst einem zusätzlichen follow-up als notwendig angesehen, um Aussagen über die Wirksamkeit einer Evaluation treffen zu können.

Der Unterschied der quasi-experimentellen zu experimentellen Versuchsplänen besteht darin, dass die Experimental- und Kontrollgruppe nicht randomisiert zugeteilt werden. Daher besteht die Möglichkeit einer Konfundierung von Interventionseffekten mit Störvariablen und damit einer Gefährdung der Validität. Eine Systematik typischer Störeinflüsse der internen und externen Validität wurde von Campbell und Stanley (1963) und Cook und Campbell (1979) aufgestellt. Die interne Validität betrifft die Eindeutigkeit und die externe Validität die Generalisierbarkeit der Untersuchungsergebnisse auf andere Orte, Personen oder Situationen.

Für den gewählten quasi-experimentellen Untersuchungsansatz könnten Einschränkungen der internen und externen Validität die Aussagekraft der Ergebnisse gefährden, daher werden einzelne Einflussfaktoren etwas genauer beleuchtet. Die Übersicht folgt der Zusammenstellung von Campbell und Stanley (1963), die hier jedoch nur in Ausschnitten wiedergegeben wird. Die interne Validität ist nach Campbell und Stanley (1963) gefährdet, wenn Effekte nicht eindeutig auf die Intervention, sondern auf andere rivalisierende Einflüsse zurückgeführt werden können. Folgende Einflussfaktoren werden genannt:

- Externe zeitliche Einflüsse: Andere als die untersuchten Einflussfaktoren haben die Veränderung bewirkt, beispielsweise wenn Aufklärungskampagnen in den öffentlichen Medien zur Sensibilisierung gegenüber der Wahrnehmung einer AIDSbedrohung führen.

- Reifungsprozesse: Die Untersuchungsteilnehmer selbst ändern sich und durchlaufen altersgemäße Entwicklungsprozesse unabhängig von der Untersuchung. Das Erlangen der Geschlechtsreife im Verlaufe einer Intervention dürfte auch die psychosoziale Entwicklung entscheidend stimulieren und Einstellungsänderungen bezüglich sexueller Inhalte bewirken.

- Testübung: Das Untersuchungsinstrument beeinflusst das zu Messende, z.B. allein durch das Ausfüllen eines Einstellungstests werden zu messende Einstellungen verändert. Eine Möglichkeit der Kontrolle der Pretesteffekte ist die Anwendung des Solomon-Vier-Gruppen-Plans (vgl. z.B. Kvalem, Sundet, Rivo, Eilertsen & Bakketeig, 1996). Kvalem et al. (1996) stellten eine Konfundierung der Programmwirkung eines schulischen Sexualaufklärungsprogramms bezüglich des Gebrauchs von Kondomen mit dem Pretest fest. Pretest oder Sexualaufklärung alleine konnten keine Zunahme des Kondomgebrauchs bei Schülerinnen und Schülern bewirken.

- Mangelnde instrumentelle Reliabilität wird in der Evaluationsforschung, wo häufig nicht genügend Zeit für die Instrumentenentwicklung bleibt als eine der Ursachen für den

fehlenden Nachweis von Programmeffekten gesehen. Die Auswertung von Differenzwerten soll jedoch laut Hager (1998) bezüglich ihrer Reliabilität bezüglich gruppenstatistischer Auswertungen unproblematisch sein.

- Statistische Regressionseffekte: Bei einer quasi-experimentellen Untersuchung zur Überprüfung von Veränderungshypothesen besteht die Gefahr, dass die Ergebnisse durch sogenannte Regressionseffekte verfälscht werden. Hiermit sind statistische Artefakte gemeint, die durch mangelnde Reliabilität der Messinstrumente verursacht sind. Extreme Pretestwerte haben die Tendenz, sich bei einer wiederholen Messung zur Mitte der Merkmalsverteilung hin zu verändern.

- Selektionseffekte: Vor allem wenn die Gruppen quasi-experimentell gebildet wurden, können Gruppenunterschiede resultieren, die mit dem selektiven Zugang zu den Gruppen und nicht durch die Maßnahme selbst erklärbar sind, die jedoch mit der Intervention interagieren. So ist nicht auszuschließen, dass innovative Modellprogramme häufiger in Schulen realisiert werden können, die durch ein gutes Schulklima und ausgeprägtes Interesse an Gesundheitsförderungsmaßnahmen charakterisiert sind. Das eigentliche forschungstechnische Problem besteht hier darin, eine adäquate Kontrollschule mit vergleichbarem Klima zu finden.

- Experimentelle Mortalität: Es kommt zu einem systematischen Ausfall von Teilnehmern in einer der Gruppen, was sich auf die Ergebnisse verfälschend auswirkt. Eine Möglichkeit, das Ausmaß an Verzerrungen abzuschätzen besteht darin, die Programmteilnehmer mit Abbrechern zu vergleichen (vgl. Mittag & Hager, 1998).

Zu den Gefährdungen der externen Validität zählen die folgenden Einflussgrößen:

- Mangelnde instrumentelle Validität: Das Messinstrument erfasst nicht das, was es eigentlich erfassen sollte.

- Stichprobenfehler: Untersuchungsergebnisse einer Stichprobe dürfen nicht auf Grundgesamtheiten generalisiert werden, für die die Stichprobe nicht repräsentativ ist.

- Experimentelle Reaktivität: Ergebnisse sind nur unter den Bedingungen gültig, unter denen sie ermittelt wurden, Generalisierungen reaktiver Messungen sind generell problematisch.

- Hawthorne-Effekte: Das Bewusstsein, Teilnehmer einer wissenschaftlichen Untersuchung oder eines Interventionsprogramms zu sein, verändert das Verhalten. Im allgemeinen führt dies bei den Betroffenen zu dem Gefühl besonderer Zuwendung und daher zu einer unspezifischen positiven Wirkung, die unabhängig von den konkreten Inhalten der Intervention ist. Eine Kontrolle der programmunspezifischen Zuwendungswirkung kann nur durch eine Kontrollgruppe mit Treatment erfolgen, die jedoch den spezifischen Programmeffekt nicht hervorrufen soll (Hager, 1998).

Campbell und Stanley (1963) gingen davon aus, dass bei einer ausreichenden Sicherung der internen Validität durch Randomisierung, Konstanthalten oder Standardisierung die beobachteten Effekte eindeutig interpretiert werden können und kausale Schlussfolgerungen zulässig sind. Leider ist der Ertrag der quantitativ-experimentellen Richtung in der Evaluationsforschung sowohl die Effekte, als auch deren Replizierbarkeit betreffend, enttäuschend gering geblieben.

7.3. Wissenschaftliches Vorgehen und Praxisforschung - ein Gegensatz?

Als einflussreicher Vertreter einer Gegenposition zu Campbell, Stanley und Cook gilt Cronbach (1982). Cronbach ist ebenso wie Campbell und Cook der quantitativ nomothetischen Richtung in der Evaluationsforschung zuzuordnen (Wittmann, 1990). Allerdings favorisierte er die korrelationsanalytische Methode, die in Differentieller und Allgemeiner Psychologie dominant ist. Diese wurde jedoch gerade von experimentell ausgerichteten Forschern nicht akzeptiert.

Cronbach (1982) kritisiert zunächst das Campbell'sche Konzept der internen Validität, das er für die praxisbezogene Forschung als wenig brauchbar ansieht. Nach Cronbach's Ansicht lässt sich Campbell's Ideal kausaler Aussagen in der Feldforschung nicht einlösen, da Kausalität im Campbell'schen Sinne immer nur in Bezug auf eine klar definierte Untersuchungssituation, mit einem bestimmten Treatment, das unter bestimmten Randbedingungen zu einem bestimmten Zeitpunkt durchgeführt wurde, gegeben sein kann. Da Campbell und Stanley die Randbedingungen oder das Setting, innerhalb dessen Interventionen ihre Wirkung entfalten, zu wenig berücksichtigen, ja diese als bloße Störeinflüsse betrachten, vernachlässigen sie den Aspekt der Generalisierung der Untersuchungsergebnisse zugunsten der von ihnen favorisierten internen Validität. Nach Cronbach steht jedoch die Reproduzierbarkeit der Wirkung einer Interventionsmaßnahme und damit die Übertragbarkeit auf verschiedene Settings im Vordergrund. Cronbach schrieb der Generalisierbarkeit der Untersuchungsergebnisse die höchste Bedeutung zu, was jedoch von einigen Autoren durchaus kritisch gesehen wird (Brandtstädter, 1990; Cook & Matt, 1990).

Auch Brandtstädter (1990) stellt ein zu starkes Festhalten am Konzept der internen Validität bei der wissenschaftlichen Bewertung von Interventions- und Reformprojekten in Frage. Brandtstädter weist vor allem auf das Problem hin, dass ein Design, das wissenschaftlichen Standards genügt, noch nicht ausreicht, um Befunde eindeutig zu erklären. Entscheidend ist vielmehr, in welchem Ausmaß theoretisch begründbare Alternativerklärungen zur Erklärung der Befunde vorhanden sind. Kann ein Untersuchungsergebnis potentiell durch viele konkurrierende Einflussgrößen zustande gekommen sein, die zudem über eine hohe Plausibilität verfügen, dann bleibt ein Experiment in seiner Aussagekraft schwach. In diesem Sinne kann auch eine Untersuchung mit einem „schwachen" quasiexperimentellen Design hohe interne Validität besitzen, wenn Beobachtungsbefunde nur auf eine Weise erklärt werden können.

Die methodische Kunst der Evaluationsforschung besteht nach Brandtstädter (1990, S. 217) darin, „auch unter feldexperimentellen Bedingungen zu hinreichend validen bedingungsanalytischen Folgerungen zu gelangen". Schlussfolgerungen, die Wirksamkeit von Interventionsmaßnahmen betreffend, sind demnach vor allem dann erlaubt, wenn kaum plausible Alternativerklärungen vorhanden sind.

Ob nun ein Untersuchungsergebnis über bedingungsanalytische Schlussfolgerungen hinausgehend auch generalisierbar ist oder nicht, ist jedoch unabhängig von dem gewählten Design. Generalisierbarkeit setzt vielmehr die Feststellung der „Äquivalenz" der Untersuchungssituation mit derjenigen Situation, auf die die Ergebnisse hin übertragen

werden sollen, voraus. Äquivalenzfeststellungen erfordern eine theoriegeleitete Unterscheidung des Untersuchungskontextes in Merkmale, die ergebnisrelevant sind und solche, die nicht relevant sind.

Auch Brandtstädter weist also ebenso wie Cronbach dem Untersuchungskontext (Setting) eine wesentliche Bedeutung in der Evaluationsforschung zu. Cronbach vertritt den Standpunkt, dass Programmwirkungen im Grunde genommen nie durch eine Intervention allein, sondern vielmehr nur durch das Zusammenwirken von Treatment und Setting zustande kommen. Entsprechend werden bei Cronbach den Interaktionen zwischen dem Untersuchungskontext und dem eigentlichen Treatment eine sehr viel höhere Bedeutung zugewiesen, als dies innerhalb des Campbell'schen Ansatzes der Fall ist, der im wesentlichen nur an einfachen Haupteffekten interessiert ist. Weiterhin war Cronbach auch an Interaktionen zwischen experimentell manipulierten Instruktionsmethoden mit differentiellen Eigenschaften der Lernenden interessiert. Cronbach setzte dies im sogenannten Aptitude Treatment Interaction Forschungsparadigma (ATI) um (vgl. Cronbach & Snow, 1977).

Indem Cronbach die Komplexität der Rahmenbedingungen praktischer Interventionen als konstitutive Bedingungen der Intervention selbst berücksichtigt, sind verbesserte Aussagen über die Anwendbarkeit der Forschungsresultate zugelassen, also Voraussagen darüber, unter welchen Bedingungen sich Interventionen als wirksam erweisen werden. Damit unterstützt Cronbach den Fokus praxisbezogener Evaluationsforscher (z.B. von Kardoff, 1993; Patton, 1982). In der Evaluationsterminologie wird die Programmwirkung unter Alltagsbedingungen als „Effectiveness" bezeichnet, die Feststellung unter idealen Modellprojektwirkungen als „Efficacy" (z.B. Walter & Schwartz, 1997). Unter Alltagsbedingungen werden im allgemeinen geringere Programmwirkungen als unter idealen Modellprojektwirkungen erwartet.

7.4. Rollenverständnis und Dilemma der Evaluationsforscher

Die Dominanz des quantitativ – experimentellen Vorgehens und der Sicherung der internen Validität in der Evaluationsforschung wurde häufig von solchen Autoren kritisiert, die den praktischen Nutzen und die unmittelbare Verwertbarkeit der Evaluationsergebnisse für den Auftraggeber als wichtigstes Ziel ihrer Arbeit ansehen (vgl. Patton, 1982; von Kardoff, 1993). Cronbach (1982, S. 321-339) vertritt beispielsweise die Ansicht, dass Evaluation eine Kunst des Möglichen ist, die sich pragmatischen Kriterien unterzuordnen habe, wenn sie dem Auftraggeber bzw. Projektträger verständliche und nützliche Entscheidungsgrundlagen beschaffen möchte. Von seiten der Auftraggeber wurde tatsächlich auch die mangelnde Praxisnähe der Untersuchungen kritisiert. Häufig wurde es dem Auftraggeber überlassen, aus der Berichtlegung eigene Schlussfolgerungen zu ziehen.

An die Rolle der Evaluatoren werden von verschiedener Seite Ansprüche und Erwartungen gerichtet, die nicht immer miteinander vereinbar sind und daher zu Rollenkonflikten bei den beteiligten Wissenschaftlern führen können. Während die Auftraggeber in erster Linie an praktisch und „strategisch" verwertbaren Informationen

interessiert sind, besteht gleichzeitig die Erwartung, dass die gewonnenen Erkenntnisse theoretisch fundiert und wissenschaftlich gesichert seien und als unangreifbare Argumentationsbasis für politische Ziele dienen können. Die „Doppelzielsetzung“ der Begleitforschung besteht hierbei darin, einerseits relevante Entscheidungshilfen für ein bestehendes Problem, das auf einen konkreten Kontext bezogen ist, zu erlangen, andererseits aber auch „gültige, zuverlässige und verallgemeinerbare Erkenntnisse über die untersuchten Handlungssysteme zu liefern“ (vgl. Ehrlich, 1995, S. 32).

Die kritisierte mangelnde Nutzenorientierung mag unter anderem auch darin begründet sein, dass häufig die Zielsetzungen des Auftraggebers vor Beginn der Evaluation nicht geklärt sind (Schmalohr, 1989). Die Unterstützung von Zielfindungsprozessen des Auftraggebers kann als Teil des Auftrags des Evaluators definiert werden, widerspricht jedoch dem vorherrschenden Rollenbild des Evaluators als Grundlagenforscher (Wottawa, 1991). Dieses Rollenbild impliziert eine Orientierung am Instrumentarium der empirischen Projektforschung.

Im Gegensatz zu dieser müssen allerdings im Bereich der Feldforschung häufig Abstriche gemacht werden. Zum einen ist in den allermeisten Fällen eine randomisierte Stichprobenziehung nicht möglich. Zum zweiten sind Einschränkungen bezüglich der Versuchsplanung hinzunehmen. Insbesondere bei großen und praxisrelevanten Feldstudien gelingt es häufig nicht, wenigstens die wichtigsten Faktoren der zu evaluierenden Maßnahme systematisch zu kombinieren. Rein quantitativ ergebnisorientierte Evaluationsforscher sehen sich daher in einem Dilemma. Sie werden auf der einen Seite von den Grundlagenforschern, deren Methoden sie übernommen haben, aufgrund unzureichender wissenschaftlicher Standards kritisiert. Andererseits sind aber ihre Aussagen, die sich an Falsifikationskonzepten orientieren, für die Auftraggeber als Entscheidungsgrundlage nicht aussagekräftig genug. Diese benötigen „positiv wirkende Gestaltungshilfen, fundierte Aussagen über die Größe von Effekten und ein möglichst umfassendes Anregungspotential für die praxisrelevante Entscheidungsfindung“ (Wottawa, 1991, S. 165). Aber auch eine allzu starke Ausrichtung an Nützlichkeitserwägungen kann gefährlich sein, wenn sie sich auf einen Utilitarismus stützt, der den historischen Kontext des zu evaluierenden Programms völlig ausblendet. Klotter (1997) fordert hier, dass Evaluation um eine gesellschaftliche Analyse ergänzt werden muss.

7.5. Die Bedeutung qualitativer Methoden für die Evaluationsforschung

Die einseitige Konzentration auf den Nachweis von Programmwirkungen und gleichzeitige Vernachlässigung formativer Evaluation brachte es zwangsläufig mit sich, dass man zuwenig darüber wusste, wie Programmerfolge oder Misserfolge überhaupt zustande gekommen waren. Eine solche einseitige Schwerpunktsetzung vermittelt gar den Eindruck, als ob die Qualität der Durchführung von Interventionsmaßnahmen keine notwendige Bedingung für den Erfolg von Interventionsprogrammen darstellen würde. In der neueren Evaluationsforschung wird dem Prozess der Durchführung der Interventionsmaßnahmen größere Aufmerksamkeit geschenkt. Der Vorteil prozessorientierter Evaluation liegt nicht

nur darin, wesentliche Bedingungsfaktoren für den Programmerfolg zu identifizieren, sondern auch in dem Potential, wichtige Informationen für die Programmentwicklung zu liefern und so als Modell für zukünftige Programmentwicklung zu dienen.

Parallel zu dieser Entwicklung wird das Primat quantitativer Forschung zunehmend in Frage gestellt und – vor allem im englischsprachigen Bereich – durch qualitative Verfahren ergänzt (Walter & Schwartz, 1997). Zu Beginn rief die Anwendung qualitativer Methoden jedoch Widerstände hervor. Typischerweise wurde argumentiert, dass die Zielsetzungen beider Methoden so unterschiedlich seien, dass eine Vermischung weder möglich noch wünschenswert sei (Rosenberg, 1988). Tatsächlich liegt der Fokus innerhalb des quantitativen Paradigmas beim Feststellen von Kausalitäten beziehungsweise Wirkungen. In der qualitativen Forschung steht demgegenüber das Verständnis und die Bedeutung sozialer Phänomene im Vordergrund. Qualitative Forscher beanspruchen eine höhere Validität der meist durch Interviews erhobenen „Daten". Es scheint, als ob für die heutige Evaluationsforschung die bekannten Ressentiments und Gegensätze der quantitativen und qualitativen Forschung nicht mehr gelten. Die Einbeziehung subjektiver Sichtweisen in die vorliegende Evaluationsstudie anerkennt die Tatsache, dass verschiedene am Programm beteiligte Gruppen und Parteien ihre Programmerfahrungen unterschiedlich rekonstruieren. Die Notwendigkeit einer Einbeziehung qualitativer Methoden in der vorliegenden Arbeit ergibt sich nicht zuletzt durch das „niedrige" Entwicklungsalter der Programm - Zielgruppe. Seit den 70er Jahren ist man von der ausschließlichen Verwendung quantitativer Methoden in der Jugendforschung abgekommen (z.B Held, 1989; Krüger, 1989). Es wird als fraglich angesehen, ob die Übertragung von vorgefassten Konzepten, die von erwachsenen Experten entwickelt wurden, der subjektiven Realität Jugendlicher gerecht werden kann.

Die meisten Evaluationsforscher sind sich heute darüber einig, dass Methodenvielfalt vor einseitigem methodischem Zugang zu bevorzugen ist. „Ein dogmatischer experimenteller Rigorismus ist zumal bei der Evaluation komplexer Reform- und Interventionsprojekte, wo sich explorative mit hypothesenprüfenden Problemstellungen mischen, ebenso unangemessen wie ein einseitig qualitativer Impressionismus." (Brandtstädter, 1990, S.219). Die Methodenauswahl geschieht daher je nach Sachlage. Experimentelles Vorgehen sollte somit ergänzt werden durch nichtexperimentelle multivariate Techniken, Fallstudien und qualitative Analysen.

Verschiedene Möglichkeiten des Zusammenwirkens qualitativer und quantitativer Methoden in der Evaluationsforschung sind denkbar. Typischerweise werden qualitative Daten zur Exploration eines unbekannten Gegenstandsbereichs eingesetzt. Aus der Perspektive des quantitativen Forschers werden sie als Vorstufe des hypothesengeleiteten quantitativen Forschens betrachtet. Oder sie werden genutzt, um quantitative Messinstrumente zu entwickeln. Eine solche Sichtweise qualitativer Forschung ordnet diese der quantitativen Forschung unter. Steckler und Koautoren (Steckler, McLeroy, Goodman, Bird & McCormick, 1992) haben außer diesem Modell (Modell 1) noch drei weitere Möglichkeiten des Zusammenwirkens qualitativer und quantitativer Daten in der Evaluationsforschung beschrieben. So können qualitativ erhobene Informationen der Erklärung quantitativer Befunde dienen (Modell 2). Dies entspricht beispielsweise dem

Einsatz qualitativer Instrumente in der Evaluationsforschung mit dem Ziel der Prozessevaluation. Umgekehrt können auch quantitativ erhobene Ergebnisse genutzt werden, um eine primär qualitative Studie besser interpretieren zu können (Modell 3). Schließlich können beide Methoden parallel genutzt werden, um Ergebnisse wechselseitig zu validieren. Beispielsweise werden Programmwirkungen sowohl mit einem geschlossenen Fragebogen, als auch mit einem offenen Interview erfragt (Modell 4). Aus der Perspektive der qualitativen Methode wird der Vergleich mit der quantitativen Methode als Triangulation betrachtet. Es wird klar, dass beide Methoden nicht konkurrieren, sondern sich wechselseitig ergänzen und vervollständigen. In der vorliegenden Evaluationsstudie waren insbesondere Modell 2 und Modell 4 von Bedeutung.

Zu ergänzen ist, dass Prozessevaluation nicht mit qualitativer Forschung gleichgesetzt werden soll. Vielmehr können sowohl quantitative als auch qualitative Methoden der Prozessevaluation dienen. Qualitative Methoden ermöglichen allerdings ein eher strukturoffenes Vorgehen im Vergleich zu strukturtestendem Vorgehen. Kromrey (1995) weist auf die Besonderheit der Modellprogramme hin, die sich in einem Praxisfeld bewegen, in dem es an theoretischen und an Erfahrungswissen mangelt. Daher ist das zu Beginn formulierte Wirkungsmodell ebenso wie das darauf zugeschnittene Forschungsdesign veränderungsoffen anzulegen.

7.6. Probleme der Bestimmung von Zielen in der Evaluationsforschung

Um eine Evaluierung zu ermöglichen, müssen Ziele vorgegeben werden, die durch das Programm verwirklicht werden sollen. Im allgemeinen werden diese Ziele nicht von den Evaluatoren, sondern von den Auftraggebern oder Praktikern vorgegeben. Diese sind zu Beginn jedoch häufig in einer diffusen und allgemeinen Sprache gehalten, die zur empirischen Überprüfung nicht geeignet sind. Die Aufgabe der Evaluatoren besteht hier darin, gemeinsam mit Planern, Projektleitern und Trägern die allgemeinen Ziele in präzise, klar definierte und konsistente Aussagen (operationale Ziele) zu verwandeln. Diese stellen Erfolgskriterien zur Bewertung des untersuchten Programms dar. Zielkriterien dürfen sich aber nicht einseitig auf gewünschte Wirkungen beziehen, sondern sollten auch unerwünschte Nebenwirkungen und nichtintendierte Wirkungen erfassen. Letzteres impliziert, dass Programmziele nicht isoliert, sondern innerhalb eines umfassenderen Kontextes zu betrachten sind. Die Auswahl der Erfolgskriterien sollte sehr sorgfältig erfolgen, da sie letzten Endes darüber entscheidet, ob vorhandene Programmwirkungen durch die Evaluation nachweisbar sind oder nicht.

Walker und Avis (1999) sehen als wesentliche Ursache für das Fehlschlagen von Peer-Education-Programmen einen Mangel an allgemeinen und operationalen Zielen für das Projekt. Dieser Kritikpunkt betrifft in besonderem Ausmaß gering strukturierte Programme. Gefordert wird, dass die Ziele realistisch und messbar sein sollten. So sollte beispielsweise bei AIDS Peer-Education-Programmen nicht die Reduktion von HIV-Infektion als übergeordnetes Globalziel formuliert werden, sondern als konkretes

Verhaltensziel das Aufsuchen von Beratungsstellen oder Gesundheitsdiensten. Erst so ist eine angemessene Evaluation überhaupt möglich und können generalisierende Aussagen über die Wirksamkeit des Programms getroffen werden.

Ein ganz anderes Problem ist die mit Zielfindungsprozessen implizit verbundene Entscheidung über die Wertigkeit von Zielen. Pädagogische Zielsetzungen sind nicht selten auch mit pädagogischen Wertvorstellungen verbunden, denen ein bestimmtes Menschenbild zugrunde liegt. Beispielsweise sieht Wottawa (1991) als eine Besonderheit der Evaluationsforschung im pädagogischen Bereich ihre ideologische Verbundenheit mit Zielen. Nicht selten existieren miteinander konkurrierende Auffassungen über adäquate Zielsetzungen. So unterscheidet sich das Programmziel „sexuelle Entscheidungsfähigkeit fördern“ vom Programmziel „sexuelle Abstinenz“ nicht nur, sondern steht sogar im Widerspruch zu dem erstgenannten. Solche Widersprüche aufzudecken und einer Klärung zuzuführen gehört mit zu den Aufgaben der Evaluationsforschung, nicht zuletzt da das Fehlen einer einheitlichen Zielsetzung die Evaluation erschwert. Zielsetzungen von seiten der Auftraggeber lassen sich zwei Polen zuordnen. Sind diese akkomodativ, das heißt werden die Programmbedingungen hilfreich an die Zielgruppe angepasst ? Oder aber wird umgekehrt die Zielgruppe an die Programmbedingungen angepasst - in diesem Falle ist die Zielsetzung als assimilativ zu charakterisieren. Eine zielfreie Evaluation – unabhängig von den Interessen der Auftraggeber - muss jedoch als unrealistisch bewertet werden.

Ein völlige Unabhängigkeit der Tätigkeit der Evaluationsforscher von Wertentscheidungen scheint ebenfalls unrealistisch. Die Evaluationsforschung ist gezwungen, Entscheidungen im Angesicht relativer Unsicherheit zu treffen. Dabei muss sie behutsam vorgehen und ihre Bewertungen in ein System von Werthierarchien einbetten und mit Blick auf situative und kontextuelle Bedingungen vornehmen.

7.7. Probleme der Implementation pädagogischer Programme

Die besonderen Schwierigkeiten der Implementation und Evaluation pädagogischer Programme innerhalb der Schule wurden z.B. von Meyer, Miller und Herman (1993) diskutiert. Das Ausmaß der Widerstände gegenüber einem Programm kann variieren von direkter Ablehnung bis hin zum Ignorieren eines Programms und mag von der Kontaktaufnahme und adäquaten Informierung der beteiligten Lehrkräfte möglicherweise aber auch von vorhandenen Werten und Einstellungen abhängig sein. So erklärten beispielsweise Weed und Jensen (1993) sowie DeGaston, Jensen, Weed und Tanas (1994) den mangelnden Erfolg eines Sexualaufklärungsprogramms durch die mangelnde Unterstützung des Programmziels „sexuelle Abstinenz“ seitens der Lehrer. Offensichtlich bestand ein enger Zusammenhang zwischen der Einstellung der Lehrer zu den Programmzielen und der Güte der Implementation des Programms anhand vorgegebener Manuale.

Die Frage, wie viel Zeit für die Durchführung von Veranstaltungen und Befragungen zur Verfügung gestellt wird, ist ganz wesentlich von dem aktuellen Lehrplan abhängig. Für die Akzeptanz des Programms durch die beteiligten Lehrkräfte mag es wichtig sein, solche

schuljahresbedingten Engpässe im Auge zu haben und den Beginn der Durchführungsphase auf eine eher belastungsfreie Zeit zu legen. Der Beginn des Schulhalbjahres ist dafür besser als das Schuljahresende geeignet. Eine Einschränkung besteht auch hinsichtlich der zur Verfügung stehenden Zeitperiode für Veranstaltungen, die rigide an den Schulstunden orientiert sein muss, um den schulischen Ablauf nicht zu behindern.

Auch Walker und Avis (1999) sehen die Passung von Programm und Setting als zentrales Kriterium für das Gelingen der Implementation von Peer-Education-Programmen. Als Setting wird häufig die Schule gewählt, nicht selten jedoch auch Freizeiteinrichtungen. Mögliche Inkonsistenzen oder durch das Umfeld bedingte Restriktionen sollten bei der Projektplanung durchaus mitbedacht werden, denn nicht alle Projekttypen sind für jedes Setting gleichermaßen geeignet. So mag ein eher strukturiertes Vorgehen eher im schulischen Kontext sinnvoll sein, hingegen ein unstrukturiertes Vorgehen im Jugendfreizeitbereich angebracht sein. Wenn das angestrebte Projektdesign nicht mit den Erfordernissen der beteiligten Organisation in Einklang gebracht werden kann, dann wird die Implementation des Programms gehemmt.

7.8. Evaluationskonzept der vorliegenden Studie

Die vorgegebene Zielsetzung, Aussagen zur Programmwirksamkeit eines Peer-Education-Programms treffen zu können, gab den grundsätzlichen Rahmen für das Forschungsdesign vor - bevorzugt wurde ein quasi-experimentelles Design mit Kontrollgruppe. Die Erfolgskriterien wurden anhand von quantitativen Methoden gemessen, in Teilbereichen war eine Validierung anhand von Interviewaussagen der Jugendlichen möglich. Qualitative Methoden wurden jedoch hauptsächlich zur Prozessevaluation eingesetzt.

Die vorliegende Programmevaluation war lose an die umschriebenen Zielsetzungen Rossis „multi-goal“ und „theory-driven“ angelehnt. So sollte der Versuch unternommen werden, auf Fragestellungen verschiedener am Programm beteiligten Interessengruppen einzugehen. Hervorzuheben ist hier aufgrund des intensiven Kontaktes mit dem Evaluationsteam die Gruppe der Praktiker. Die beteiligten Trainer widmeten als Mitarbeiter der Senatsverwaltung für Gesundheit über einen Zeitraum von dreieinhalb Jahren den größten Teil ihrer Arbeitszeit dem Projekt. Sie waren in erster Linie an Rückmeldungen interessiert, die sie in ihre praktische Arbeit einfließen lassen können. Im besten Falle konnten sie die Evaluation als ein Instrument „kontinuierlicher Qualitätssicherung“ nutzen, wenn sie Feedback aus der Prozessevaluation erhielten. Im schlechtesten Falle fühlen sie sich in ihrer Arbeit von meist „fachfremden“ Personen kontrolliert. Konflikte entstehen meist dann, wenn der Wunsch der Evaluatoren, möglichst umfassend und standardisiert Daten zu erheben, an die Grenzen der Praktikabilität stoßen und aus Sicht der Praktiker den Ablauf eines Modellprojektes stören, ohne dass für sie ein Nutzen des von ihnen geleisteten „Evaluationsaufwandes“ ersichtlich wäre. Im Verlaufe der Durchführung der Evaluationsstudie wurden sowohl die Praktiker, als auch die Peer-Educators und Schülerinnen und Schüler über die sie betreffenden Ergebnisse der

Evaluationsstudie informiert. Ein solches Vorgehen stimmt mit ethischen Standards der Evaluationsforschung überein. Eine genaue Beschreibung der Einzelergebnisse würde den Rahmen dieser Arbeit sprengen.

Als zweite Direktive sollen in Anlehnung an Rossi latente dem Programm zugrunde liegende Theorien identifiziert und im Rahmen der Evaluationsstudie überprüft werden. Damit werden auch Interessen der beteiligen Evaluationsforscher als Grundlagenforscher angesprochen. Als geeignet zur empirischen Analyse erweist sich die Theorie des Modellernens. Das Evaluationsdesign beinhaltet eine Bewertung der Charakteristika von Peer-Educators durch Adressaten und ermöglicht daher eine Analyse der wahrgenommenen Modellwirkungen in Relation zu den Programmwirkungen. Zum zweiten ist prinzipiell eine Überprüfung der gesundheitspsychologischen Theorie als Grundlage der Interventionsmaßnahme möglich. Nach Sheeran, Abraham und Orbell (1999) müssen effiziente AIDS-Präventionsprogramme auf Kognitionen abzielen, die reliabel mit dem Kondomgebrauch zusammenhängen und gleichzeitig der Prävention durch Modifikation zugänglich sind. Als erfolgversprechend haben sich hier die behavioralen und normativen Überzeugungen, Einstellungen, Selbstwirksamkeitserwartungen und Intentionen erwiesen, die Einzelkomponenten der Theory of Planned Behavior darstellen. Mit Rossi wird davon ausgegangen, dass eine Einbeziehung theoriebasierter Variablen zu einem vertieften Verständnis kausaler Beziehungen beitragen und das Vertrauen in die Generalisierbarkeit der Forschungsergebnisse erhöhen kann.

Allerdings gilt eine wichtige Einschränkung. Die wenigsten Peer-Education-Programme wurden explizit auf der Grundlage einer sozialwissenschaftlichen Theorie entwickelt. Auch wenn dies der Fall war, ist nicht davon auszugehen, dass die implementierten Programmelemente logisch stringent von solchen Theorien ableitbar sind. Prinzipiell ist eine Vielzahl von Realisierungen von Theorien in Programme denkbar.

Nach diesen Erläuterungen zu Problemen der Evaluation folgt eine Literaturübersicht über den Stand der Evaluationsforschung zu Auswirkungen von Peer-Education-Programmen auf Multiplikatoren und Adressaten.